A COMPARISON OF ALTERNATIVE VARIANCE ESTIMATORS IN DEMOGRAPHIC SURVEYS

Ophelia M. Mendoza, University of the Philippines
William D. Kalsbeek, University of North Carolina

When the need exists for computing the variance of survey estimates, the investigator is often confronted by two things which make the problem difficult: a complex estimate and a complex sampling design (see Kish and Frankel, 1974). An estimate such as a ratio or correlation coefficient is considered complex because the mathematical formula for computing it is a complex (i.e., nonlinear) function of random variables. On the other hand, stratification and cluster sampling in multiple stages are characteristics which make the sampling design complex. The difficulty arises since variance formulae for complex estimates obtained from complex sampling designs are themselves complex functions, if expressed precisely, thereby making computations burdensome.

The investigator is left with several possible approaches to variance estimation. The preferred approach, of course, is to construct a simple variance estimate which adequately addresses the complex nature of the estimate and sampling design. A simple approach would be to assume (1) that the sample had been selected by simple random sampling rather than by the more complex sampling design that was actually used and (2) that the estimate can be expressed in a simple form (e.g., as a simple function of a proportion). As somewhat of a compromise, a third approach would be to assume simple random sampling, but to allow the estimate to retain its complex form.

This paper is the outgrowth of an empirical study designed to answer the question: to what extent do variances computed by the simple and compromise approaches differ from variances produced by the preferred approach? We consider the specific setting in which survey data from a complex sample of households in Indonesia are used to produce various complex indirect estimates of childhood q(x) survivorship attributable to Sullivan (1972) and Trussell (1974).

Ours is not the first attempt to obtain variance estimates for demographic measures. Earlier studies by Chiang (1960), Keyfitz (1966), O'Brien (1981), and Retherford and Bennett (1977) have considered the problem of producing variance estimates for the expectation of life, the net reproduction rate, several life table survivorship functions, and the own-children method for obtaining indirect estimates of fertility, respectively. Aside from the O'Brien paper, ours is the first known attempt to contrast alternative variance estimates.

As one final preliminary note, it must be understood that the variance estimates produced in this study and presented in the tables take into account only the sampling error of measures taken directly from the survey. They do not accommodate those modelling errors which result from producing the coefficients of the multiplicative adjustment factor for q(x). The ideal complete measure of error for q(x) would be one in which both modelling and sampling errors are jointly considered. This, to our knowledge, has not yet been attempted and may be pointed to as the subject of subsequent research.

INDIRECT MORTALITY ESTIMATORS

In response to known deficiencies in vital registration systems and population censuses, various indirect methods for estimating fertility and mortality have been developed. We consider in the present study indirect methods due to Sullivan (1972) and Trussell (1974) for estimating the q(x) survivorship measures of early childhood. Both methods follow the procedure originally proposed by Brass (1968) of converting, by means of a multiplicative factor, statistics on the proportion dead of children ever born to women in certain five-year age intervals.

The Sullivan estimator utilizes a multiplicative factor which is a simple linear function of the estimated ratio of the average parity of women aged 20-24 to women aged 25-29. This parity ratio was found to be highly correlated with the age at onset of childbearing and is easily obtained from the data necessary for the calculation of the age-specific proportion dead among children ever born to women. The estimator for the probability of dying between birth and exact age x is given by

$$q(x) = D_i[A_i^{(S)} + B_i^{(S)}(P_2/P_3)] \quad , \qquad (1)$$

where

D_i = estimated total number of children deceased among those ever born to women in the i-th age group / estimated total number of children ever born to women in the i-th age group,

P_2 = estimated total number of children ever born to women aged 20-24 / estimated total number of women aged 20-24,

P_3 = estimated total number of children ever born to women aged 25-29 / estimated total number of women aged 25-29,

$A_i^{(S)}$, $B_i^{(S)}$ = constant coefficients obtained external to the survey. The convention we adopt for the subscript, i, is that i=1 refers to the age group 15-19, i=2 refers to the age group 20-24, and so forth. For x=2, 3 and 5 considered in this study, i=2, 3 and 4 respectively.

The Trussell estimator uses a slightly more complex multiplicative factor which is a linear function of the estimated parity ratio for women aged 15-19 and 20-24 and of the estimated parity ratio used in the Sullivan estimator. Here the estimator is

$$q(x) = D_i[A_i^{(T)} + B_i^{(T)}(P_1/P_2) + C_i^{(T)}(P_2/P_3)] \quad , \quad (2)$$

where $A_i^{(T)}$, $B_i^{(T)}$, $C_i^{(T)}$ = constant coefficients obexternal to the survey. The relationship between

values of x and i are the same as with the Sullivan estimate.

The reason for considering both the Sullivan and Trussell estimators in this study is that they differ in degree of complexity. The Trussell estimator, as we shall see in the next section, is a function of a larger number of random variables than the Sullivan estimator. Given this difference in complexity, it will be of interest to note the relationship between the complexity of an estimate and corresponding variance estimates.

METHOD

In this section we describe three alternative estimates one might conceivably use in producing the variance of the Sullivan and Trussell indirect estimates from complex sampling designs. The first estimate, even though it uses an approximation, is most technically appropriate since the nature of the estimator and the sampling design are both properly accommodated. We call this the preferred variance estimate. The second variance estimate takes into account the nature of the indirect estimate but ignores the actual sampling design by assuming simple random sampling instead. We call this the compromise variance estimate. The third variance estimate assumes a simpler form of the estimate and simple random sampling. We call this the simple variance estimate.

Preferred Variance Estimate

The Taylor series linearization (TSL) or delta method of producing variance estimate was adapted for sample surveys by Woodruff (1971) using a procedure discussed in Chiang (1968). Suppose that the mathematical formula for the indirect estimate takes the general form $\theta = \phi(\underset{\sim}{t})$, where $\phi(\underset{\sim}{t})$ is a nonlinear function of $\underset{\sim}{t} = (t_1, t_2, \ldots, t_g)$, a g-dimensional vector of estimated totals. The measure being estimated is $\phi(\underset{\sim}{T})$, where $\underset{\sim}{T} = (T_1, T_2, \ldots, T_g)$ is the corresponding g-dimensional vector of actual totals. We note from the terms of (1) that g = 5 or 6 for the Sullivan estimator and from the terms of (2) that g = 7 or 8 for the Trussell estimator.

The estimate θ is produced from a stratified multistage cluster sample in which the symbol, h, indexes one of H primary strata; the symbol, α, denotes one of a_h PSU's in the h-th stratum, and the symbol, β, is used to index one of the $b_{h\alpha}$ elementary sampling units selected in the α-th PSU of the h-th primary stratum. Using data from this design, we obtain the j-th estimated total of $\underset{\sim}{t}$ as

$$t_j = \sum_{h}^{H} \sum_{\alpha} \sum_{\beta}^{a_h b_{h\alpha}} W_{h\alpha\beta} t_{jh\alpha\beta} , \qquad (3)$$

where $W_{h\alpha\beta}$ is the sampling weight defined as the reciprocal of the selection probability for the β-th sample elementary unit in the α-th PSU of the h-th stratum, and $t_{jh\alpha\beta}$ is the measurement on the $h\alpha\beta$-th elementary unit used to produce t_j.

Using the first-order terms of a Taylor series expansion of $\phi(\underset{\sim}{t})$ about $\phi(\underset{\sim}{T})$ and following the general approach to variance estimation suggested by Hansen, Hurwitz, and Madow (1953), the preferred variance estimate for $\theta = \phi(\underset{\sim}{t})$ is calculated as

$$\mathrm{var}_p(\theta) = \sum_{h}^{H} \left\{ \frac{a_h \sum_{\alpha} z_{h\alpha}^2 - \left(\sum_{\alpha} z_{h\alpha}\right)^2}{a_h - 1} \right\} \qquad (4)$$

where

$$z_{h\alpha} = \sum_{\beta}^{b_{h\alpha}} z^*_{h\alpha\beta} = \sum_{\beta}^{b_{h\alpha}} W_{h\alpha\beta} z_{h\alpha\beta} , \qquad (5)$$

$$z_{h\alpha\beta} = \sum_{j}^{g} \phi_j^{(1)}(\underset{\sim}{t})\, t_{jh\alpha\beta} . \qquad (6)$$

and $\phi_j^{(1)}(\underset{\sim}{t})$ is the first partial derivative of $\phi(\underset{\sim}{t})$ with respect to t_j.

Compromise Variance Estimate

The preferred variance estimate in (4) is applicable to a stratified multistage cluster sample. If we were to consider the complex nature of θ but assume that simple random sampling had been used instead, the resulting compromise variance estimate, ignoring the negligible finite population correction, would be

$$\mathrm{var}_c(\theta) = n s_z^2 , \qquad (7)$$

where $n = \sum_{h}^{H} \sum_{\alpha}^{a_h} b_{h\alpha}$ is the overall sample size and

$$s_z^2 = \frac{n \sum_{h}^{H} \sum_{\alpha}^{a_h} \sum_{\beta}^{b_{h\alpha}} z^{*2}_{h\alpha\beta} - \left(\sum_{h}^{H} \sum_{\alpha}^{a_h} \sum_{\beta}^{b_{h\alpha}} z^*_{h\alpha\beta}\right)^2}{n(n-1)}$$

Simple Variance Estimate

Both Sullivan's and Trussell's indirect estimates can be viewed as adjusted proportions with the linear factors in (1) and (2), respectively, as the adjustments. If we were willing to assume that the linear adjustment factors in both of the estimates are constants, then the indirect estimates would take the general form, $\theta = kD_i$, where k is the assumed constant and D_i is the proportion (dead among children ever born to women in the i-th group). Using the binomial variance and the well-known fact that $\mathrm{Var}(kD_i) = k^2 \mathrm{Var}(d_i)$, the simple variance estimate, assuming that simple random sampling was used and ignoring the finite population correction once again, would be com-

puted as

$$\text{var}_s(\theta) = k^2 D_i(1 - D_i)/n_i \ , \qquad (9)$$

where n_i is the sample size, defined here as the number of children ever born to women who were selected in the sample and who fall in the i-th age group.

EAST JAVA POPULATION SURVEY

Data for this empirical study were obtained from the initial round of the East Java Population Survey (EJPS) conducted in 1980 by the Central Bureau of Statistics in Indonesia, in collaboration with the International Population Laboratories at the University of North Carolina. The nonself-weighting sample of households interviewed in the initial round was selected in three stages with desas, census blocks, and households as sampling units in the first, second, and third stages, respectively. Desas, serving as PSU's, were stratified at two levels. First, all urban desas were stratified by municipality and all rural desas were stratified by regency. Within each of these major strata, desas were ordered by population density, and approximately equal-sized zones were constructed. In each of the resulting 1,038 rural zones, one desa was selected using a form of systematic sampling with probability roughly proportional to the number of households. Using a similar form of systematic sampling, two desas were selected independently in each of 100 urban zones. Thus, the total number of PSU's was 1,238. Each desa selected as a PSU was subdivided into so-called census blocks, mostly containing 60-75 households. One census block was then randomly selected in each sample desa. Sampling units in the third stage were households. For purposes of sample selection, a list of households was prepared for each selected census block. Households chosen for the survey were selected from these lists by systematic sampling after a random start. The final sample of 19,772 households contributed data for 19,111 ever-married women aged 15-49 that were used for this study.

To meet the requirement of two PSU's per stratum for variance estimation, pseudo rural strata were formed by combining pairs of zones following the original ordering of the first stage sampling frame. Formation of pseudo strata in urban zones was not required since two PSU's had been selected per zone.

FINDINGS

Tables 1 and 2 contain estimates of variance for the Sullivan's and Trussell's indirect estimates of the childhood survivorship measures, q(2), q(3), and q(5). A "West" mortality pattern was assumed for all survivorship estimates. All estimates are presented for male children only, female children only, and for both sexes combined.

Indirect estimates of q(x) and a comparison between the preferred and compromise variance estimates are presented in Table 1. The sample size (n) in Table 1 is the number of ever-married women who contributed data to the indirect estimate or its accompanying variance estimate. For example, Sullivan's q(2) and its accompanying variance estimate use data from women in two separate age groups, 20-24 and 25-29. Sample sizes for q(2) and q(3) are the same since the same age groups provide sample data. Table 1 also contains the ratio, $R_c = \text{var}_p(\theta)/\text{var}_c(\theta)$, which measures the relative size of the preferred and compromise variances. Defined in this manner, R_c is what Kish (1965, Section 8.2) calls deff (denoting the "design effect") which represents the variance for the complex design divided by the variance for a simple random sample of the same size.

Some of the key findings from Table 1 are now presented. First, Sullivan's and Trussell's estimates of q(x) are quite similar which is consistent with the findings of Trussell (1974). Moreover, comparable values of R_c are also similar. Second, $R_c > 1$ indicating that the compromise variance estimate is smaller than the preferred variance estimate. We note, furthermore, that $1.06 \leq R_c \leq 1.33$ which is quire low when recalling that R_c = deff, which in other demographic surveys is generally higher (see Kish, et al., 1976). The smaller values reported here are undoubtedly due to the relatively small average cluster sizes resulting from the large number of sample clusters selected in the EJPS. Third, values of R_c are generally smaller for estimates obtained for each sex separately than they are for estimates obtained for both sexes together. A direct explanation of this pattern is not possible without a better understanding of the unknown statistical properties of indirect estimates; however, it is interesting to note that the opposite pattern is observed with $\text{var}_p(\theta)$ whose values are larger for male and female estimates than for both sexes combined. Finally, we observe a modest direct relationship between R_c and g, the number of random variables used to produce q(x). It is difficult to judge whether this relationship is real or is attributable to sampling error. If it is indeed real, the trend may be due to the order of approximation in using the linear terms of the Taylor series expansion to produce $\text{var}_p(\theta)$.

The preferred and simple variances computed for q(x) are compared in Table 2. The sample size, n_i, used to compute $\text{var}_s(\theta)$ is the number of children ever born to women in the sample falling in the age group used to produce D_i for q(x). The ratio, $R_s = \text{var}_p(\theta)/\text{var}_s(\theta)$, indicates the relative difference between the preferred and simple variances.

The major findings of Table 2 are now presented. First, values of R_s are always notably larger than corresponding values of R_c indicating that the simple variance estimate is even smaller than the compromise variance estimate and is therefore substantially smaller than the preferred variance estimate. We note that $1.15 \leq R_s \leq 1.98$, indicating that the preferred variance of q(x) will be 15 to 98 percent greater than the simple variance.

Larger values of R_c, as might be expected in other surveys, would contribute to even greater values of R_s. Second, as with R_c, corresponding values of R_s are similar for Sullivan's and Trussell's estimates. Third, we observe once again that values of R_s are greater for q(x) produced for all children than for q(x) produced for males and female children separately. Fourth, values of R_s become notably larger as one proceeds from q(2) to q(3) to q(5). This difference may be partly due to the general tendency for k to decrease as we move from q(2) to q(3) to q(5). Larger values of k would contribute to larger $var_s(\theta)$ and to smaller R_s. Finally, unlike our observation for R_c in Table 1, there is no apparent direct relationship between g and R_s.

DISCUSSION

Variances presented here apply to indirect mortality estimates for all women in the EJPS sample. Similar variances were also prepared but not reported here for various subgroups of the population of ever-married women aged 15-49 (e.g., region, occupation of the head of household, woman's educational attainment, urban-rural, etc.). Trends as those discussed earlier are less evident when considering these variance estimates because sample sizes for these subgroups are relatively small so that all estimates produced from them are subject to greater sampling errors than are estimates produced for all ever-married women aged 15-49.

Viewing the results in Tables 1 and 2 collectively we conclude that, when estimating q(x) by the Sullivan or Trussell approach, the simple variance estimate tends to be substantially smaller than the preferred variance estimate. The compromise variance estimate will also be smaller than the preferred estimate but to a much lesser degree. Considering computational ease, the simple variance estimate is clearly the least difficult to produce, requiring only k and D_i (which are available in producing q(x)) and n_i (which can be easily obtained as a raw frequency from the survey data). The preferred and compromise variance estimates are about equally difficult to produce, both requiring a linearization value ($z_{h\alpha\beta}$) for each member of the sample. Computational formulae are both relatively simple functions of the linearization values. One might therefore reasonably conclude that the preferred variance estimate is best used if one has gone to the trouble of computing $z_{h\alpha\beta}$. If it is impractical to compute the $z_{h\alpha\beta}$, then the simple variance estimate can be used, but it will tend to produce rather severe underestimates.

One might be tempted to suggest that the best strategy in estimating variances for q(x) is to compute the simple variance estimate and then multiply it by some factor reflecting the likely size of R_s. The problem with this strategy is in choosing a suitable multiplicative factor. As we have seen in Table 2, values of R_s vary considerably. Moreover, the size of R_s depends on the nature of the sampling design and other factors that are not currently well-understood. Observed values of R_s might, for example, be considerably larger in sampling designs with larger average cluster sizes than EJPS. Thus, we conclude that R_s lacks adequate portability, referring to properties of R_s which allow its use far from its source (see Kish, et al., 1976). Using the simple variance estimate in combination with R_s, if done at all, is best done with extreme care.

REFERENCES

[1] Brass, W. (1968), The Demography of Tropical Africa, Princeton University Press, Princeton, N.J.

[2] Chiang, C.L. (1960), "A Stochastic Study of the Life Table and Its Applications: II. Sample Variance of the Observed Expectation of Life and Other Biometric Functions," Human Biology 12:221-238.

[3] Chiang, C.L. (1968), Introduction to Stochastic Processes, John Wiley and Sons, New York.

[4] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), Sample Survey Methods and Theory, Vol. II, John Wiley and Sons, New York.

[5] Keyfitz, N. (1966), "Sampling Variance of Demographic Characteristics," Human Biology 38:22-41.

[6] Kish, L. (1965), Survey Sampling, John Wiley and Sons, New York.

[7] Kish, L., and Frankel, M.R. (1979), "Inference from Complex Samples," Journal of the Royal Statistical Society, Series B, 36:1-37.

[8] Kish, L., Groves, R.M., and Krotki, K.P. (1976), "Sampling Errors for Fertility Surveys," WFS Occasional Paper No. 17, World Fertility Survey, London.

[9] O'Brien, K.F. (1981), "Life Table Analysis for Complex Survey Data," Institute of Statistics Mimeo Series, No. 1337, Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill.

[10] Retherford, R.D. and Bennett, N.C. (1977), "Sampling Variability of Own-Children Fertility Estimates, Demography 14:571-580.

[11] Sullivan, J.M. (1972), "Models for Estimating the Probability of Dying Between Birth and Exact Ages of Early Childhood," Population Studies, 26:79-97.

[12] Trussell, T.J. (1974), "A Re-estimation of the Multiplying Factors for the Brass Technique for Determining Childhood Survivorship Rates," Population Studies, 29:97-108.

[13] Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, 66:411-417.

Table 1: Comparison of Preferred and Compromise Variance Estimates of Sullivan's and Trussell's Indirect Estimates (West Mortality Varient) of q(x) for East Java, by Sex of Children

	Age (x)	Number of Variables (g)	Sample Size (n)≠	θ = q(x)			$var_p(\theta)$*			$R_c = var_p(\theta)/var_c(\theta)$			Average R_c
				Males	Females	Both Sexes	Males	Females	Both Sexes	Males	Females	Both Sexes	
Sullivan:													
	2	5	7,789	0.1382	0.1049	0.1220	0.76	0.56	0.39	1.15	1.06	1.18	1.15
	3	5	7,789	0.1359	0.1123	0.1243	0.42	0.36	0.24	1.16	1.12	1.22	
	5	6	10,539	0.1391	0.1243	0.1320	0.41	0.42	0.26	1.16	1.16	1.21	1.18
Trussell:													
	2	7	12,435	0.1394	0.1055	0.1229	0.78	0.57	0.40	1.14	1.06	1.18	1.19
	3	7	12,435	0.1382	0.1142	0.1264	0.44	0.39	0.27	1.21	1.20	1.33	
	5	8	15,185	0.1435	0.1283	0.1362	0.43	0.45	0.28	1.14	1.18	1.22	1.18

≠This is the number of ever-married women in the sample contributing data to the estimate.

*Estimates must be multiplied times 10^{-4}.

Table 2: Comparison of Preferred and Simple Variance Estimates of Sullivan's and Trussell's Indirect Estimates (West Mortality Varient) of q(x) for East Java, by Sex of Children

	Age (x)	Age Group (i)	Proportion dead among children ever born to women in i-th age group (D_i)			Adjustment Factor (k)			Sample Size for $var_s(\theta)$ ($n_i^{\neq}$)			$R_s = var_p(\theta)/var_s(\theta)$		
			Males	Females	Both Sexes	Males	Females	Both Sexes	Males	Females	Both Sexes	Males	Females	Both Sexes
Sullivan:														
	2	20-24	0.133	0.100	0.117	1.04	1.05	1.04	2,144	2,034	4,178	1.31	1.15	1.45
	3	25-29	0.139	0.115	0.127	0.98	0.98	0.98	3,879	3,742	7,621	1.43	1.38	1.72
	5	30-34	0.143	0.128	0.135	0.97	0.97	0.98	4,350	4,047	8,397	1.54	1.61	1.95
Trussell:														
	2	20-24	0.133	0.100	0.117	1.05	1.06	1.05	2,144	2,034	4,178	1.32	1.16	1.47
	3	25-29	0.139	0.115	0.127	0.99	0.99	1.00	3,879	3,742	7,621	1.44	1.45	1.87
	5	30-34	0.143	0.128	0.135	1.00	1.00	1.01	4,350	4,047	8,397	1.51	1.63	1.98

≠This sample size for q(2), q(3), and q(5) corresponds to the number of children ever born to women in the sample aged 20-24, 25-29, and 30-34, respectively.