Narayan Das

On the Estimation of Sampling Error in Certain Fertility Indices

SAMPLE surveys are commonly used to assess the demographic situation in ^countries where vital statistics are incomplete or absent and to assess the impact of birth control programme. But if the data are faulty and the estimates are subject to sampling error, the demographic position and changes can not be assessed unless the degree to which survey data are correct is known. The results of a sample survey are affected not only by sampling errors arising from chance variation in the selection of the sample but also by non-sampling errors such as lack of precision in reporting observations, incomplete or faulty canvassing of a designated random sample etc. (Sukhatme, 1953, p. 10), However, the present paper deals only with the sampling error.

The average magnitude of sampling errors depends on the size of the sample, on the variability of the material, on the sampling procedure adopted, and on the way in which the results are calculated (Yates, 1960, p. 2). The methods by which the average magnitude of the sampling errors can be calculated depend on the mathematical theory of sampling. The knowledge of such sampling errors in relation to the characteristic under study enables further surveys to be more efficiently planned.

The purpose of this paper is to discuss sampling error in relation to certain fertility indices and its use in the determination of the size of the sample, when independent sampling in each age group of the female population is done.

Usually the fertility indices are computed from data on births and deaths relating the total female population. Rarely, independent sampling in each age group is done to estimate such fertility rates. What normally is drawn is a random sample of size *n* from total female population, or a stratified random sample of size *n,* with proportional allocation.

The fertility indices which are considered for the present study are age specific fertility rates, total fertility rate, gross reproduction rate and net reproduction rate. To estimate these indices v\e need only be concerned with the females in the reproductive span of life usually stretching from 1 5 to 44 or 49 years; more strictly, we are concerned with the all married females in the reproductive span.

Method of Sampling

Let us consider a situation in which the sampling units in the population are the females in the reproductive span and stratified into different age groups. The sampling units under each age group can be divided into two mutually exclusive classes, class 1 consisting of females who had given a birth during the preceding year (or during the reference year) and class 2 comprising those who had not.

Let *pi* be the proportion of females of age group (i — 12, *i + 1/2* |) in the population belonging to class 1, and q_i the proportion of females of age group *(i —1/2* , i + 12) falling in class 2. If *Ni* is the total number of females of age group (i — f, *i* + 0 in the population, *N,pi* will be number of females of that] age group in the population belonging to class 1 and $N_i q_i$ the number of* females of the same age group in class 2 so that $N,pi + N_i q_i$ -- *Ni.*

In a sample of n_t females each drawn without replacement from finite population of size *Ni,* by the method of simple random sampling, let us suppose that n_{ib} females are found to be in class 1 and the rest («, — *nib)* in class 2.

In such a distribution n,b is regarded as a hyper geometric variate. Since the possible values which n,b can assume are 0, 1, 2 . . . *m.*

Hence $E(n_{tb}) = n_i p_t$ and

$$V(n_{tb}) = \frac{N_i - m}{N_i - 1} n_i p_i q_i .$$

When N_i is sufficiently large n_{ib} follows binomial distribution and variance of n_{ib} is given by $V(n_{ib}) = n_i p_i q_i$. In case of the proportion n_{ib}/n_i

$$E\left(\frac{n_{ib}}{n_i}\right) = p_i \ .$$

Hence an unbiased estimate of the age specific fertility rate (proportion p_i) in the population is given by the proportion in the sample.

$\hat{p_i} = n_{ib}/n_i = p_i'$ and its variance is given by

$$V(p_i') = V(n_{ib}/n_i) = p_i q_i/n_i = p_i(1 - p_i)/n_i.$$

Hence total fertility rate (T) which is the sum of the age specific fertility rate for women belonging to the age group 15-44 or 15-49 is given by $T = \sum_{i=15}^{49} p_i$ whose unbiased estimate will be $T' = \sum_{i=15}^{49} p_i'$.

Since the samples have been drawn independently at random from each age group of the reproductive span so that covariances are zero, it can be easily shown that the variance of T' is :

$$V(T') = \sum_{i=15}^{49} \frac{p_i(1 - p_i)}{n_i} \text{ and its estimate is :}$$

$$\hat{V}(T') = \sum_{i=15}^{49} \frac{p_i' 1 - p_i'}{n_i - 1}.$$

Similarly by treating each female as a sampling unit and assigning to each an equal probability of being selected, if f_i is the proportion of the mothers with female births in the population, and $(1 - f_i)$ is the proportion of not having female births, the variance of the proportion of mothers with female births in a random sample of n_i units can be obtained and hence the variance of the gross reproduction rate (G) and net reproduction rate (R). We have similar

results given by Koop (1951, p. 156) as

$$G' = \sum_{i=15}^{49} f_i' \text{ and } R' = \sum_{i=15}^{49} l_i f_i'$$

where G and R are the unbiased estimates of G and R respectively, 1 is the female survival rate for age / and is assumed to be constant. The variances of G' and *R'* which measure sampling errors are given by

$$V(G') = \sum_{i=15}^{49} \frac{f_i(1-f_i)}{n_i}$$

and

$$V(R') = \sum_{i=15}^{49} l_i^2 \frac{f_i(1-f_i)}{n_i}$$

respectively. Their estimates are :

$$V(G') = \sum_{i=15}^{49} \frac{f_i'(1-f_i')}{n_i - 1}$$

$$V(R') = \sum_{i=15}^{49} \frac{l_i^2 f_i'(1-f_i')}{n_i - 1}$$

respectively in case of effectively large number of females in each age group.

However, if 1* is not constant and subject to error, the variance of *R'* will include terms for variance due to fertility and mortality, and for covariance between them. If the samples for fertility and mortality have been drawn independentfy, the covariance term would drop out and similar result as given by Keyfitz (1969, p. 570) would appear :

$$V(R') = \sum_{i=15}^{49} l^2 V(f_i) + \sum f^2 V(l_i)$$

Since it is assumed here that mortality is subject to negligible error, the second term would drop out and the earlier result would appear again.

Using the same methods and assumptions as for age specific fertility rate, the estimates of sampling error associated with crude birth rate and general fertility rate could also be obtained. In the present study not all types of fertility indices are considered,

To carry the discussion further on the relative accuracy of these fertility indices, above estimates of G and R can also be obtained from the estimates of T and written as

$$G' = \sum_{i=15}^{49} B_i^f p_i' \text{ and } R' = \sum_{i=15}^{49} l_i B_i^f p_i'$$

with estimates of their variances

$$\hat{V}(G') = \sum_{i=15}^{49} (B_i^f)^2 \frac{p_i'(1-p_i')}{n_i - 1}$$

$$\hat{V}(R') = \sum_{i=15}^{49} l_i^2 (B_i^f)^2 \frac{p_i'(1-p_i')}{n_i - 1}$$

respectively, where $B^f i$ is the proportion of female births to total births of mother aged i and l_t is the survival rate for specific age i.

Since B^f and 1_1 are assumed to be constant for a specific age and $B_{\cdot} <, 1$ and $li \prec 1$ it can be shown that

$$\hat{V}(T') \geqslant \hat{V}(G') \geqslant \hat{V}(R').$$

What might be of more interest is any relationships between the estimated relvariances or estimated coefficients of variation of those fertility indices.

To get a specific idea about efficiencies of the fertility indices under consideration, it is desirable to have the percentage standard error (Yates, 1960, p. 95) of the estimate of all the fertility indices considered. The standard error of an estimate is a measure of the average magnitude of the random sampling to be expected in that estimate. It also provides an indication of the frequency with which errors of various magnitude are expected to occur. The percentage

standard error of the estimate of the total number of units having 3 given attribute in the population is the same as the percentage standard error of its estimate. From the above formula this percentage standard error (PSE) of T', G' and K is given by

$$\text{P.S.E. } (T') = \frac{\sqrt{\sum_{i=15}^{49} p_i(1-p_i)/n_i}}{\sum_{i=15}^{49}} \times 100;$$

$$\text{P.S.E. } (G') = \frac{\sqrt{\sum_{i=15}^{49} (B_i^f)^2\, p_i(1-p_i)/n_i}}{\sum_{i=15}^{49} B_i^f\, p_i} \times 100;$$

and

$$\text{P.S.E. } (R') = \frac{\sqrt{\sum_{i=15}^{49} 1_i^2\, (B_i^f)^2\, p_i\,(1-p_i)/n_i}}{\sum_{i=15}^{49} 1_i\, B_i^f\, p_i} \times 100$$

respectively, whereas for any particular age group the P.S.E. of Pi,Bi and *It B{pi* remains same while estimating 7", G' and R' respectively.

If the values of B and li are constant over the reproductive age groups, the percentage standard error of the above indices (7", G" and R') can be seen to be same. Hence the percentage standard error of 7" will differ from that of G and K depending on the variability of the values of B^f and the values of $B^f i$ and 1, over reproductive age group respectively. Similarly, the difference in the values of PSE of G' and that of R' depends on the variation of the values of U's. Since C7' is in practice less influenced by variation in the proportion of female births to total births by mother's age, and R' is also more influenced by variation in fertility than in mortality and proportion of the female to total births, the indices will have more or less the same percentage standard error.

But if the proportion of the female to total births (B) and survival ratio (10 are not constant and obtained from different survey it can be easily shown that

for **any** particular age group PSE of />,, $B\{ p<$ and h $B'i_{Pi}$ does not remain the same, while estimating T, G' and R respectively, and holds a relation of the type:

$$\text{PSE of } p_i < \text{PSE of } B_i\, p_i < \text{PSE of } 1_i\, B'_i\, p_i,$$

where pi, B^f_f and U are assumed to be independent on the basis that they are obtained from three different surveys.

Determination of Sample Size

From the above discussion, it is clear that the knowledge of PSE of either of the above mentioned fertility indices (preferably the one whose coefficient of variation is the largest) from previous survey is sufficient to estimate the sample size, when all the rates are required to be estimated from the data of a single sample. For example, in case of considering age specific fertility rate, the above formula may be rewritten so as to give the m required for the sample when the required percentage standard error of p_t is known. We have

$$n_i = \frac{10{,}000 q_i}{p_i\,(\text{required PSE})^2}$$

where/ - 15, 16 . . . 49.

Considering the previously obtained maximum and minimum PSE of the age specific fertility rates out of all age groups (between 15-49) as required for each group, minimum and maximum sample size could be obtained for each age group separately and hence for all the age group (between 15-19). Thus the variance of T' is given by

$$V(T') = \sum_{i=15}^{49} \frac{p_i(1-p_i)}{n_i}$$

such that $\frac{1}{n_0} \sum_{i=15}^{49} p_i\,(1-p_i) > V(T') > \frac{1}{n_L} \sum_{i=15}^{49} p_i\,(1-p_i)$

where «• is smallest of all », and "1 is largest of all n,.

Similarly the sample size could be estimated using any other fertility indices discussed above. These formulae hold only for a random sample in which the sampling units are the ones where the proportion with a given attribute has to be estimated.

Practicability of the Sampling System

In the case of a small population, living within a compact geographical area, the above sampling procedure may be suitable and practicable since this sampling system presupposes the availability of age composition of the population. Of course, census records can furnish such information. Because of the changes of population from births, deaths and migration, census records are not completely satisfactory unless a survey is timed to occur soon after a national census. However, selecting the sample in this way is definitely better than selecting it on the basis of a simple random sample of size « (not n^), because breaking up a sample size *n* into *m* (for different age groups) destroys the random character of the sample and it may not be possible to estimate exact sampling error of those fertility indices based on n*.

In the case of a large population of, say, millions, spread over a very wide area, the application of the above sampling procedure will be very difficult if not entirely impracticable. The alternative procedure as suggested by Koop (1957, pp. 162-165) is that the area over which the population is distributed is divided into a number of grids in such a way that each grid could be controlled easily while locating, listing and stratifying (in yearly age groups) all the females in the reproductive age groups. Since the total area is supposed to be very large, the number of grids will be very large and each will contain a large number of persons. Therefore, a random sample of the grids is selected and in each grid the variates *m* and ,•&, as defined already are observed. Then simply summing over *m* and n_{ib} for particular age group over the grid and taking the ratio of two ($«_{i6}$/n,-) an estimate of the index (ASFR) along with the estimate of variance could be derived,

Usual Sampling System of Estimating Fertility Indices

The above mentioned fertility indices are usually computed from data on births and deaths relating to the total female population. Thus the estimates are always based on a simple random sample of size *n,* or a stratified random sample of size, », with proportional allocation. If independent sampling in

tach age group is done and n is sufficiently large, the chances of these fertility indices suffering from certain biases are small. But rarely, independent sampling in each age group is done to estimate these rates by methods of statistical sampling, which help in estimating their accuracy.

It is in this context, as an alternative method, that an attempt is made to estimate these fertility indices (TFR, GRR and NRR) along with their variances, when a simple random sample of size n is drawn from the total female population,

To arrive at this type of indices along with their sampling errors, let us consider a situation in which the sampling units in the population are females in the reproductive age span.

In a random sample of size n drawn without replacement from a finite population of size N, let us suppose that m females are in the ith age group and *tin,* females are of age i with experience of births during the preceeding year. The *HI* and rtiB are regarded as binomial variates when the population is large. Hence anunbiasedestimateoftheproportion$p(i)orp(i,b)$inthepopulationisgivenbytheproportionii by the proportion in the sample.

TABLE 1

Age	*Birth During the Reference Year*					
	Population			*Sample*		
	Given	*Not given*	*Total*	*Given*	*Not given*	*Total*
i-th age group	N_{ib}	$N_i - N_{ib}$	N_i	n_{ib}	$n_i - n_{ib}$	n_i
Not in the *i*-th age group	$N_b - N_{ib}$	$(N - N_b) - (N_i - N_{ib})$	$N - N_i$	$n_b - n_{ib}$	$(n - n_b) - (n_i - n_{ib})$	$n - n_i$
	N_b	$N - N_b$	N	n_b	$n - n_b$	n

$\hat{p}(i) = n_i/n = p'(i)$ and its variance is given by

$$V\{p'(i)\} = V(n_i/n) = p(i)\{1 - p(i)\}/n.$$

Similarly,

$$\hat{p}(i, b) = n_{ib}/n = p'(i, b) \text{ and its variance is given by}$$

$$V\{p'(i, b)\} = V(n_{ib}/n) = p(i, b)\{1 - p(i, b)\}/n.$$

Therefore for **a** given age i the estimate of age specific fertility rate, *p(bji)* is **given by**

$$p'(b/i) = \frac{p'(i, b)}{p'(i)} = \frac{n_{ib}}{n_i}$$

where P (i) is the probability that a woman selected at random is of age / and *p* (i, b)* is the probability that a woman selected at random is found to be mother of age *i* who had a birth during the preceeding year.

Now, the problem for consideration is the derivation of the expected value and the variance of the sample proportion *nn,lm.* Since both the numerater and the denominator are random variables, the expected value is given by

$$E\left(\frac{n_{ib}}{n_i}\right) = E\left[E\left\{\frac{n_{ib}}{n_i}\middle/ n_i\right\}\right]$$

where *E* j — - /Hi > denotes the conditional expectation of «,&/'«* for given m and the second *E* denotes the expected value of the function in *m* so obtained.

From the earlier discussion, we know that *niblm* follows a hyper-geometric distribution in sample of ni drawn from *Ni.* Hence

$$E\left\{\frac{n_{ib}}{n_i}\middle/ n_i\right\} = \frac{N_{ib}}{N_i} = \frac{p(i, b)}{p(i)}$$

where

$$p(i, b) = \frac{N_{ib}}{N} \quad \text{and } p(i) = \frac{N_i}{N}.$$

Substituting the above result, it is clear that

$$E\left(\frac{n_{ib}}{n_i}\right) = E\left\{\frac{N_{ib}}{N_i}\right\} \simeq \frac{N_{ib}}{N_i} = \frac{p(i,b)}{p(i)}.$$

Hence an unbiased estimate of the age specific fertility rate, *p(bji)* in the population is given by the age fertility rate,/)'(6/i) in the sample.

To obtain the variance of *rtu/nt = p (b/i* we proceed similarly (, 1953, pp. 51-54) and have the results as

$$V_1\left(\frac{n_{ib}}{n_i}\right) = \frac{1}{np(i)} \frac{p(i,b)}{p(i)}\left\{1 - \frac{p(i,b)}{p(i)}\right\}$$

and $$V_2\left(\frac{n_{ib}}{n_i}\right) = \frac{1}{np(i)} \frac{p(i,b)}{p(i)}\left\{1 - \frac{p(i,b)}{p(i)}\right\}\left[1 + \frac{1}{np(i)}\{1 - p(i)\}\right]$$

where *V1* and V_3 are the first and second approximations to the variance respectively and population is assumed to be large for simplification.

Therefore, it is possible to estimate age specific fertility rates along with their approximate sampling errors, when a simple random sample of size » (irrespective of age of the women) is drawn from the total female population and hence the total fertility rate along with its approximate variance, by just summing over *i.* Similarly, considering the proportion of female births to total births and survival rate of the females for specific ages, the estimates of GRR and NRR along with their approximate sampling error could be derived. It may be pointed out, in connection with the relative accuracy of the fertility indices, that the same relation and inequality, as obtained in case of independent sampling in each age group, hold true when a random sample of size *n* is drawn from the total female population and the estimates of those indices are obtained.

The results hold for simple random sampling; however, even for other probability sampling designs, these would give approximate sample size or approximate sampling error.

Adjustment for Intraclass Correlation

As mentioned earlier all sampling error computations in this paper are based on the assumption of simple random sampling. However, usually cost and administrative considerations imply some sort of clustering in sampling design. After clustering, the selected areas may be either exhaustively enumerated or sampled again to choose a specified number of elements from each selected

cluster. In either case, such clustering causes an increase in sampling error due to effects of p, intraclass correlation (United Nations, 1971, p. 55). Usually, persons of the same cluster will have similar characteristics than those belonging to different clusters. Consequently, the actual sampling variance based on pluster,sampling will ordinarily exceed that based on the same sample size under assumptions of simple random sampling. Therefore, the formulae derived in the previous sections need be modified to take into account of the fact that demographic sample surveys are usually based on a cluster sample design.

If the average number of sample selected in each cluster is $\bar{m}$, $(\bar{m} - 1)p$ will measure the relative change in sampling variance when passing from a simple random design to a cluster sample (Sukhatme, 1953, p. 247). In the case of sub-sampling, the relative change in variance is approximately given by an expression similar to that in the case of equal or unequal cluster, viz. $(\bar{m} - 1)\,p$, where m is the number of second-stage units to be drawn from each selected first stage unit. This implies an increase in the sample size required for a given error.

It may be,noted that the intraclass correlation coefficient is apt to be relatively low (but greater than zero) for fertility measures such as the crude birth rate; or age-specific fertility rate based on births in the last 12 months but rather high for true cohort or life time fertility measures, such as number of children ever born. However, no attempt is made in this paper to estimate the sampling error associated with the latter. The estimates of cumulative fertility have substantially, smaller sampling errors than the estimates "of current fertility, because they incorporate many more women-years of experience.

References

1. Keyfitz, Nathan, 1969, Sampling for demographic variables. In : N. L. Johnson and Harry Smith, Jr.. (eds.),New' *Developments in-Survey Sampling,* Wiley-Inter Science, A Div. of John Wiley and Sons, New York, p. 569.
2. Koop, J. C., 1951, Notes on the estimation of gross and net reproduction rates by methods of statistical sampling, *Biometrics,* **VII,** 15S-166.
3. Sukhatme, P, V., 1953, *Sampling Theory of Surveys with Applications,* The Indian Society of Agricultural Statistics, New Delhi, India and Iowa State University Press, Ames, **Iowa, U.S. A.**
4. United Nations, 1971, Methodology *of. Demographic Sample Surveys,* Series M, No. 51, New York. .United Nations pp. 46-57..
5. Yates F., 1960, *Sampling Methods for Census and Surveys,* Charles Griffin and Company Limited, London, Third Edition.