ON A METHOD OF ESTIMATING BIRTH AND DEATH RATES AND THE EXTENT OF REGISTRATION

C. Chandra Sekar and W. Edwards Deming

Reprinted from the Journal of the American Statistical Association
March, 1949, Vol. 44, pp. 101-115

Reprinted from the JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
March, 1949, Vol. 44, pp. 101-115

ON A METHOD OF ESTIMATING BIRTH AND DEATH RATES AND THE EXTENT OF REGISTRATION

C. CHANDRA SEKAR
All-India Institute of Hygiene and Public Health, Calcutta
(on loan to the Population Division of the United Nations)

AND

W. EDWARDS DEMING
Bureau of the Budget, Washington

A MATHEMATICAL THEORY is presented which when applied to a comparison of the registrar's list of births and deaths with a list obtained in a house-to-house canvass, gives an estimate of the total number of events over an area in a specified period; also the extent of registration.

In the development of the theory, allowance is made for the fact that the chance of an event being missed on one list (registrar's list or the house-to-house canvass) may not be independent of its chance of being missed on the other list. Where there is likely to be lack of independence, a test is suggested and a method introduced to reduce the effect of dependence. This is done by subdividing the data into small homogeneous groups, such as might be formed by small areas, sex and age classes, domiciliary and institutional births; then by estimating the number of events in these groups separately and summing them for a total. The standard errors of the estimates are given.

The theory is applied to an enquiry that was conducted in February 1947 over an area known as the Singur Health Centre, near Calcutta, covering the years 1945 and 1946 separately, and it is found that the estimated total number of events for the area is usually greater when the estimate is built up by summing the totals for individual groups than when it is computed at once for the aggregated population. According to the theory this observation confirms positive dependence and indicates that the greater figure is nearer the truth.

The annual number of births and deaths in the Singur Health Centre (total population 64,000) is estimated subject to a standard error of from 1 to 3 per cent, and the registration is estimated to vary from about 40 to 70 per cent with a standard error of about 3 per cent. This enquiry provides basic ground work for the design of future surveys, and it is estimated that at a cost of Rs. 10,000 to Rs. 15,000 (3 rupees to the U.S. dollar) estimates of birth and death rates for an entire District in India with a population of one to two millions can be obtained with an overall standard error of about 5 per cent.

Purpose. The purpose here is to present a theory by which when vital registration is incomplete, an enquiry in the form of a house-to-house canvass may be used in conjunction with the registrar's list to estimate, *i.* the total number of births and deaths in an area over a specified period; *ii.* the birth and death rates; *iii.* the deficiencies in registration; and *iv.* the standard errors of all these estimates. The theory will first be presented, then applied to particular surveys in the Singur Health Centre.

Method of enquiry. The application of the theory which is to be developed requires a comparison of the entries on:

1. The registrar's list (referred to as R)

2. The result of a complete house-to-house canvass carried out by an interviewer (referred to as I) and the classification of the entries on these lists into the following four exhaustive groups:

C, the number of entries recorded in I and also in R (such entries, being found on both lists, are assumed to be correct without investigation).

N_1, entries recorded only in R but not in I, and after investigation found to be correct.

N_2, entries recorded only in I but not in R, and after investigation found to be correct.

X, entries recorded on one list or the other, but not both, and found after investigation to be incorrect.

This is a complete classification of the entries on the lists but not of the events. There will also be a number Y of events which are missed by both lists; this number will be estimated later by application of the theory.

Theory. Let N be the total number of events (births or deaths) in the specified period. Then an estimate $\hat{N}$ of N is furnished by the formula $\hat{N}=C+N_1+N_2+N_1N_2/C$ wherein N_1N_2/C is an estimate of the number of events Y missed by both R and I. This formula of estimation assumes that the chance of an event being missed on either list is independent of the chance of being missed on the other. A method is presented later on for investigating the validity of the assumption of independence and for introducing a modification where necessary.

It can be shown that: *i.* $\hat{N}$ is an unbiased estimate in the limit when N becomes large and the assumption just mentioned is valid; *ii.* the maximum likelihood estimate is equal to $\hat{N}$ in the limit; *iii.* the standard error of N is $\sqrt{Nq_1q_2/p_1p_2}$. The last formula will be developed in the appendix. Here,

$$p_1 = \text{the chance of } R \text{ detecting an event}$$

$$p_2 = \text{the chance of } I \text{ detecting an event}$$

$$p_1 + q_1 = p_2 + q_2 = 1.$$

It follows that the better the performance in *either* R or I, the higher be p_1 or p_2, the smaller be q_1 or q_2 and the more precise be the estimate $\hat{N}$ of the total of events. It follows, moreover, that the precision of $\hat{N}$, expressed as a proportion (namely as a coefficient of variation), is $\sqrt{q_1q_2/Np_1p_2}$, wherefore if the theory be applied over an area large enough to contain a large number N of events, the total number N of events will be estimated with great relative precision.

The symbol p_1 is a measure of performance of the registrar, an estimate for which is $\hat{p}_1 = C/(C+N_2)$. This estimate $\hat{p}_1$ of p_1 is subject to a coefficient of variation of

$$\sqrt{\frac{q_1}{(C+N_2)p_1}\cdot\frac{N-C-N_2}{N-1}}\,.$$

This error decreases as $C+N_2$ increases. For perfect performance on the part of the interviewer, $C+N_2=N$, and there is then no error in estimating the performance of the registrar.

The foregoing development is oversimplified. In practice there are some problems to take account of—incomplete investigation of the R lists; incomplete coverage of the population in the house-to-house canvass. Special types of events, like those occurring in institutions, are best taken care of as a separate group. Then again there is the problem of investigating the assumption mentioned above, and of measuring and correcting for the correlation between the chance of an event being missed by R and being missed by I. These points will be examined in the following paragraphs.

Effect of incomplete investigation of the registrar's lists. In the investigation of the R-lists there may be some entries left over unclassified by reason of incompleteness of entry, illegibility, or simple failure for any reason whatever on the part of the investigator to finish his job. So long as the correct entries amongst the unclassified entries on the R-list constitute unbiased samples from the two categories C and N_1 mentioned earlier, the omission of the unclassified entries from the calculations does not affect the estimation of N, the total number of events. The estimate of the extent of registration will be too low if the unclassified entries contain, as is likely, correct entries classifiable as C. If the unclassified entries are all counted as correct, contrary to fact, the calculations will lead to an overestimate of the extent of registration.

Effect of incompleteness of coverage of the population. As in every population enquiry, there will be some failures to elicit information from all the households. This will happen when some households in which an event took place have moved away temporarily or permanently, or when no responsible person can be found at home to give the information. So long as the events in the uninterviewed portion of households are included in the R-list to the same extent as those in the interviewed households, the estimation of N is unaffected. The calculation of N may therefore be little affected by incompleteness of coverage of the population.

The effect of institutional events. In rural areas the bulk of the births are domiciliary, but there are some small scattered hospitals drawing patients from a wide area, and a high proportion of the events that take place in them are for non-residents. The R-list may contain some or even all of the entries for these institutional events because the registrar is able to ascertain this information easily and accurately from the institutions. The interviewer, on the other hand, will, by the nature of a house-to-house canvass, fail to discover an institutional event concerning people who had no family connections in the area. Institutional events, as they are accurately ascertainable, are best handled as a separate block and not as a problem of estimation.

The effect of correlation between events missed on both lists. The first step is to define this correlation. The registrar and his co-workers will detect some events and miss others. The probability that the interviewer [I] will detect an event that was missed by R may be different from the probability that he will detect an event that was recorded by R. If these two probabilities are equal there is complete independence, but otherwise there is not, in which case the formula given above for the estimation of the total number of events will be incorrect. The extent of the error can be investigated. If as before,

p_1 = the probability of the registrar detecting an event
q_1 = the probability of the registrar missing it

then the probabilities in the 4 groups will be shown by the accompanying table, which defines four new probabilities, P_{21}, P_{22}, Q_{21}, Q_{22}. p and q are always complementary: $p_{21}+q_{21}=p_{22}+q_{22}=1$.

Group	Probability
C Detected by both	p_1p_{21}
N_1 Detected by registrar only	p_1q_{21}
N_2 Detected by interviewer only	q_1p_{22}
Y Missed by both	q_1q_{22}

If there is complete independence between the events missed by both R and I, then $p_{21}=p_{22}=p_2$, introduced previously, and $q_{21}=q_{22}=q_2$. When there is dependence the expected value of the estimate of the number of events Y missed by both R and I will be close to

$$\frac{Np_1q_{21}q_1p_{22}}{p_1p_{21}}$$

whereas the correct value is Nq_1q_{22}. The difference is

$$\frac{Np_1q_{21}q_1p_{22}}{p_1p_{21}} - Nq_1q_{22} = Nq_1\left(\frac{p_{22}}{p_{21}} - 1\right).$$

So if $p_{21}>p_{22}$, the total number of events is underestimated and if $p_{21}<p_{22}$, the converse. We surmise that $p_{21}>p_{22}$ is likely to be the case.

Similarly, in the case of dependance, the registrar's performance is estimated as $p_1p_{21}/(p_1p_{21}+q_1p_{22})$ instead of p_1, the difference being $(p_{21}-p_{22})p_1q_1/(p_1p_{21}+q_1p_{22})$. If $p_{21}>p_{22}$ the registrar's performance is overestimated and if $p_{21}<p_{22}$, the converse.

If

$$\begin{aligned} p_1 &= .8 & q_1 &= .2 \\ p_{21} &= .6 & q_{21} &= .4 \\ p_{22} &= .4 & q_{22} &= .6 \end{aligned}$$

the bias in the estimation of the total number of events will be

$$q_1\left(\frac{p_{22}}{p_{21}} - 1\right) = -.067 \text{ or } -6.7 \text{ per cent.}$$

This bias may be much more important than the standard error of an estimate of the total number of events made under the assumption of zero correlation.

Method to reduce the effect of correlation. It is important to note that correlation signifies heterogeneity in the population for it implies that events that fail to be detected do not form a random sample from the whole population of events. This heterogeneity may arise only if there are differences in the reporting rates for different segments of the population, resulting in the group of failures being weighted disproportionately by the different segments.

It therefore follows that the correlation can be minimized by dividing the population into homogeneous groups and calculating the total number of events separately for each group; then by addition getting the grand total. In order to put this suggestion into practice, let us consider the difference between two estimates of the total number of events: *i.* by dividing the population into homogeneous groups and estimating the events in each group separately, then forming a grand total; *ii.* by treating the entire population as a unit. Let the population be comprised of k homogeneous groups, with N_i events in the i-th group ($i=1, 2, \cdots, k$). Then let $p_1^{(i)}$ be the probability of the registrar detecting an event in the i-th group, and $p_2^{(i)}$ the corresponding probability for the interviewer. The expected value of the number of events missed by both in the i-th group is $N_i q_1^{(i)} q_2^{(i)}$ and for the entire population the total missed by both will be $\Sigma N_i q_1^{(i)} q_2^{(i)}$. As by definition there are only k homogeneous groups, this value will be estimated without bias when the groups are treated separately. But if the entire population of events were pooled, the expected value for the estimate of the number of events missed by both would be close to

$$\frac{\left[\sum N_i p_1^{(i)} q_2^{(i)}\right]\left[\sum N_i q_1^{(i)} p_2^{(i)}\right]}{\sum N_i p_1^{(i)} p_2^{(i)}}.$$

The difference in the two values will be

$$\frac{\left[\sum N_i p_1^{(i)} q_2^{(i)}\right]\left[\sum N_i q_1^{(i)} p_2^{(i)}\right]}{\sum N_i p_1^{(i)} p_2^{(i)}} - \sum N_i q_1^{(i)} q_2^{(i)} = -\frac{N^2 S_1 S_2 r}{\sum N_i p_1^{(i)} p_2^{(i)}}$$

where

$$S_1^2 = \frac{\sum N_i\left[p_1^{(i)} - \bar{p}_1\right]^2}{\sum N_i}$$

$$S_2^2 = \frac{\sum N_i\left[p_2^{(i)} - \bar{p}_2\right]^2}{\sum N_i}$$

$$N = \sum N_i \qquad \bar{p}_1 = \frac{\sum N_i p_1^{(i)}}{\sum N_i} \qquad \bar{p}_2 = \frac{\sum N_i p_2^{(i)}}{\sum N_i}$$

and

$$r = \frac{S_{12}}{S_1 S_2} = \frac{\sum N_i [p_1^{(i)} - \bar{p}_1][p_2^{(i)} - \bar{p}_2]}{S_1 S_2 \sum N_i}$$

is the correlation coefficient between $p_1^{(i)}$ and $p_2^{(i)}$, weighted by N_i, the number of events to which they have reference. If $r > 0$, then treating the entire population as a unit, we are led to an underestimation of the number of events missed by both parties and therefore an underestimation of the total number of events. This also results in an overestimate of the extent of registration. If this is the case, the population need be divided only to the stage when further division shows no increase in the total number of events. It should be possible by actual trial with some real data to decide whether (e.g., in computing number of deaths) 5-year age groups are a more effective subdivision than 10-year age groups; and whether infant deaths should be treated separately.

The enquiry in Singur Health Centre. The Singur Health Centre consists of four contiguous Union Boards,[1] viz., Singur, Balarambati, Bora, and Begumpur, situated in the Serampore sub-division of the Hooghly district. The village Singur which serves as the headquarters is only 21 miles away from Calcutta and is easily accessible by rail from Calcutta. The total area of the Centre is about 33 square miles and comprised of 68 villages with a total population of about 64,000 distributed over 12,000 families living in about 8,300 houses. As is usual in West Bengal, the villagers live close together in a compact block and wide fields separate such blocks. Since 1944 this area has formed the controlled practice field of the All India Institute of Hygiene and Public Health, Calcutta, for their experiment in Public Health Methodology.

Procedure for registration. The procedure for the registration of births and deaths in this area follows closely the method adopted in other parts of Bengal. The Chowkidar, i.e., the village headman, is the reporting agent and is required to submit periodically to the Sanitary Inspector,[2] who is the registrar of the area a list of births and deaths.

[1] The Bengal Province is divided into divisions, the divisions into districts, the districts into subdivisions, the subdivisions into thanas, and the thanas into Union Boards.

[2] A Sanitary Inspector is usually in charge of the health activities of a thana.

TABLE I

THE INVESTIGATORS' REPORT ON THE COMPARISON OF THE R AND I LISTS OF SINGUR HEALTH CENTRE

Event	Year	R: Registrars' lists						I: Interviewers' Lists
			Number Verified					
		Total number of events in the lists	Total	Common C found in the the inter-viewer's lists	Extra N_1 not found in Inter-viewer's lists	No. non-verifiable. illegible, incomplete etc.	Number incorrect	Extra in interviewer's lists N_2
Births listed as occurring in the village (excluding non-resident institutional)	1945	1,748	1,504	794	710	156	88	741
	1946	2,659	2,242	1,506	736	228	189	1,009
Deaths listed as occurring in the village (excluding non-resident institutional)	1945	1,356	1,083	350	733	190	83	372
	1946	1,052	866	439	427	117	69	421

With a view to improving the registration in this area, the voluntary services of a villager have been enlisted. He is not only expected to assist the *Chowkidar*, who may be illiterate, by making entries in the *Chowkidar's* register, but also to inform the registrar directly on all births and deaths in the village. The registrar also obtains a list of births, maternal and infant deaths as known to the Maternity and Child Welfare Department, and by co-ordinating the information from the three sources is expected to improve birth and death registration. For all practical purposes the voluntary agency began operating only from January 1946.

Method of enquiry. The enquiry in the Singur Centre covering 1945 and 1946 was started on the 17th February 1947. The field work lasted for eleven weeks. In this enquiry an interviewer called on every household to enumerate the resident population (separately as present and absent) and visitors with particulars of community, age, sex, and marital status, and to list all births and deaths which occurred in the village during 1945 and 1946, listing separately with relevant particulars those that occurred outside the Singur Health Centre. The lists so prepared are the I-list which, as was mentioned earlier, were compared with the registration books (the R-list). In the field-organization as actually employed, there were four investigators who worked at the comparisons and supervised the work of the 16 interviewers. The interviewers and the investigators were selected from the village population as it was thought that they would be able to obtain better cooperation than an outsider.

It should be emphasized that the comparison of the two lists is crucial. The establishment of the identity of two entries, one on one list and one on the other, sometimes requires extreme perseverance. In some cases the registrar's entry is by hearsay, and part of it may be wrong, and often much consultation is required. The interviewer's entry, however, is fortunately accompanied by a house-number or other means of identification by which the information may be verified if necessary.

Basic data obtained from the enquiry. Table I shows the results of the investigators' comparisons of the R and I-lists. As mentioned earlier, there are some problems arising from illegible and incomplete entries, the movements of the population and institutional births. The table gives some idea of the magnitude of these problems. For example the non-verifiable entries on the registrars' lists run to roughly 10% or more of their total entries. In view of their magnitude the assumption that the unverifiable entries are a representative sample of all entries, an assumption that will be made in the calculations, becomes all the more important. The need of more careful registration in the future is apparent.

No separate account was maintained of the number of correctly registered events occurring in families that had migrated out of the village prior to the interviewers' survey. The assumption will be made that the registrars would have recorded this category to the same degree as for the non-migrants, but the number is small and under the conditions of the Indian village, this assumption is not important.

In this enquiry the non-resident institutional births and deaths are considered separately and excluded from the table, as indicated. Institutional facilities exist only in the Singur Union Board. The number of the institutional births to non-residents was about 8% in 1945 and 1946. The number of institutional deaths of non-residents was only about 3%.

Estimation of total births and deaths. In order to investigate the homogeneity of smaller groups comprising the whole, so as to arrive at the best estimate of the total number of events, calculations were carried out—

i. for the Centre as a whole (births and deaths)

ii. for each Union Board separately; then these figures were combined (births and deaths)

iii. for males and females separately for the Centre as a whole; then these figures were combined (deaths only)

iv. for age groups by sex for the Centre as a whole; then the figures were combined (deaths only)

In 1945 the total number of deaths as estimated by these four methods were 2234, 2238, 2245, and 2418 respectively each with a standard error of approximately 70. In 1946 the number of deaths as estimated by the four methods were 1,696, 1,684, 1,698 and 1,765, each with a standard error of approximately 40. The closeness of the first three estimates indicates that the chances of the registrar and the interviewer detecting an event did not vary to any marked extent between Union Boards and the sexes. The increase obtained by the fourth method clearly indicates that the chances of the interviewer and the registrar detecting a death may differ considerably with the age of the dead person. Positive correlation is indicated.

Higher percentages of deaths in the younger age-groups were missed by both R and I as compared with adult age groups. The proportion missed also show a tendency to increase in the more advanced age-groups. It would be interesting to ascertain whether the estimate could be increased still further by finer subdivision of age groups or by subdivision in regard to other characteristics within each group, but no further analyses were conducted.

As for births, the total number estimated from the data of the entire Centre was 2908 for 1945 and 3744 for 1946. Separate estimation for the four Union Boards when totalled yields 2915 and 3775 for the same years. It is to be noted that while the latter figures are the higher of the two, the figure for 1945 is higher by only 1/7th of the standard error and the figure for 1946 is higher by a whole standard error.

The highest figure obtained by breaking the population into groups in various ways, and adding the estimated number of events, is to be accepted as nearest the true figure. The nonresident institutional events, which were left out of consideration may be added in to get the total number of events occurring in the area.

Estimation of rates and incompleteness of registration. For computing birth and death rates over an area, the population base is furnished by the house-to-house canvass. The total number of correct entries in the R-list judged against the total estimated number of events, measures the extent of registration. Tables II and III show the results obtained for rates and for completeness of registration.

TABLE II

BIRTH AND DEATH RATES IN 1945 AND 1946, SINGUR HEALTH CENTRE

	1945		1946	
	Rate	Standard error	Rate	Standard error
Birth rate per 1,000 population	46.1	0.8	59.8	1.0
Death rate per 1,000 population	37.7	1.2	27.5	0.7
Specific death rate (males)	36.4	1.6	27.3	1.0
Specific death rate (females)	39.2	2.1	27.8	1.0

TABLE III

PERCENTAGE OF BIRTH AND DEATH REGISTRATION DURING 1945 AND 1946

Union board	Birth registration		Death registration	
	1945	1946	1945	1946
Singur	60.4–67.9	70.9–77.1	38.1–46.9	42.0–49.1
Balarambati	51.5–55.8	53.3–57.8	45.8–55.9	50.8–58.0
Bora	53.1–61.3	56.0–66.0	54.9–66.5	52.6–63.4
Begumpur	47.4–50.3	61.3–64.7	42.6–46.4	44.9–48.1

Note (1) The range is due to non-verified entries on the *R*-list.

Note (2) The figures are subject to a standard error of about 3 per cent.

One comment may be made in regard to the birth rate for 1946, which appears to be very high. Possible explanation may be the improved economic situation after the famine of 1943, and demobilization. Another possible explanation is failure of the investigator to establish the identity of entries in the R and I lists, but if this were so, it should be more apparent for 1945, which it is not, as the birth rates for 1945 are much lower. An improbable explanation is that each Union Board is composed of extremely heterogeneous sections displaying negative correlation between the probabilities of detection of events by the Registrar and the interviewers.

Another comment should be made. The completeness of registration, recorded in Table III, is based on the number of *correct* entries in the R-list judged against the estimated total number of events. Official published rates in all countries are based on the total number of registrars' entries, correct plus incorrect, and the usual practice of inflating official rates to correct for incompleteness of registration yields spurious results: the rates are already partly inflated owing to incorrect entries. Proper inflation (correction of rates) is possible only by comparing the registration lists with the results of a population survey and making estimates of the total number of events and the proportion of incorrect entries in the registration lists.

The precision of estimated number of events. From the fact that the coefficient of variation of a total estimated number of events is $\sqrt{q_1q_2/Np_1p_2}$, it will be seen that the lower the efficiency of detection of an event on either the I or R-lists (p_1 or p_2), the greater the standard error of the total. In this enquiry, in spite of the fact that local people were hired and trained especially for this work, the efficiency of the interviewing was not of high order: only 67.2% of the births in 1946 and 52.8% in 1946 were detected by the interviewers. The corresponding percentages for deaths were 50.7 and 32.3. Methods of improving the performance of the interviewers must be sought, and it appears that the interval of time to be covered by the survey must not extend too far back.

It is highly important to bear in mind that regardless of the interviewers' performance, the method proposed here for estimating the total number of, N, events is not subject to bias,[3] but poor performance does increase the error of the estimate of N. It also increases the standard error of the estimate of the registrars' performance.

The coefficient of variation is also influenced by N. It is important to note that N in the formula refers to any total—not just a total over

[3] In making this statement the case of p_2 (or p_1) $=0$ is considered trivial and is excluded.

an area, but a total for any subgroup, such as an age or sex classes for which an estimate is prepared. For the area and sex classes that were used here, the standard errors of the estimated totals varied from 1 to 10%. Over a larger area, or over broader classes, the coefficients of variation would be reduced by the presence of the factor $\sqrt{N}$ in the denominator.

Costs. A few words regarding the cost of this particular enquiry may be helpful in planning future enquiries. The cost of the field-work, including salaries and overhead charges, amounted to Rs. 4,000. The cost of tabulation and analysis amounted to Rs. 1,500. The total cost was thus Rs. 5,500 or about $1\frac{1}{3}$ annas (2 U. S. cents) per capita in the area of enquiry. For various reasons (this being a pilot study and a complete listing of the population being desirable for other reasons), the entire population was covered without the introduction of sampling. In designing an enquiry for a larger area such as a province or even a district, sampling would be used.

For each area in the sample there can be calculated the total number of events and the rate: also the efficiency p_1 of the registrar's performance. For each sample-area, supposedly completely canvassed (no sub-sampling) there will be an error in estimating either the rate or the registrar's performance. The coefficient of variation in the rate will be the expression already given earlier, viz. $\sqrt{q_1q_2/Np_1p_2}$. Likewise, the coefficient of variation of the estimate of p_1, the registrar's performance, is

$$\sqrt{\frac{q_1}{(C+N_2)p_1} \cdot \frac{N-C-N_2}{N-1}}.$$

Each symbol refers to the particular area covered. These errors are not erased by taking a complete canvass. (As a matter of fact, the particular enquiry described here was a complete canvass, yet subject to these errors.)

When sampling is introduced to study a whole District, the estimation of the total number of events, the rates, and the over-all efficiency of registration will be made by combining the data from a number of sample-areas. An additional error is then introduced for a District as a whole because of variability between the sample areas. The variability between the rates of the individual sample-areas may be much smaller than the variability between their total events, as it is usually difficult to define sample-areas of equal populations. It follows that usually a much smaller sample will provide a standard error of (e.g.) 4 per cent in an over-all **rat**e for a District than is required to provide the same precision in the total number of events.

The cost of attaining (e.g.) a 4% error of sampling will depend on the particular design of sample that is used; and the design in turn will, for greatest economy, depend on the density and distribution of the population, on the variability of the birth and death rates over the area for which estimates are to be prepared, on the costs of purchasing or preparing maps and lists by which the sampling procedure may be formulated, on the quality of personnel available to carry out the work, etc.

As a general principle, applicable to large populations, so far as the errors of sampling are concerned, the total number of cases (i.e., the total number of people, households, areas, or whatever unit constitutes the elements of sampling) to be included in the survey depends almost entirely on the precision of sampling that is desired in the estimation of the total number of events, or in the rate (whichever is the aim of the survey) and hardly at all *on the total number of inhabitants in the area to be covered.*[4]

In India, the birth and death rates should be estimated at least by the District (roughly 1 to 2 million inhabitants), and for smaller areas if funds would permit. Roughly speaking, to attain an over-all standard error of 5% (a reasonable aim for the present), the cost of a survey will run between Rs. 10,000 and 15,000 for a district.

Additional information provided. A survey of this type also provides valuable ancillary information regarding other characteristics of the population such as size of family, age and sex distribution, marital status, occupation and industry, specific fertility rates, gross and net reproduction rates, and other information, but the list cannot be extended indefinitely because the interest of the field workers must not be dissipated too far from the main aims of the survey.

APPENDIX

THE STANDARD ERROR OF $\hat{N}$

An approximate value for the standard error of $\hat{N}$.

$$\hat{N} = \frac{(C + N_1)(C + N_2)}{C}$$

[4] It is presumed in this statement that the physical facilities for sampling (maps, lists, personnel, payment, etc.) are about the same over all parts of the area to be covered.

[5] As a matter of fact, the surveys reported have provided most of these additional items, and the cost mentioned includes them.

can be obtained by the application of the formula that the variance $Vf(x)$ of a function $f(x)$ of x is approximately given by

$$Vf(x) \simeq \left(\frac{\partial f}{\partial x}\right)_E^2 V(x)$$

where $(\;)_E$ denotes the substitution of the expected values for x appearing inside the bracket after differentiation, and $V(x)$ denotes the variance of x.

If $C+N_1$, $C+N_2$ and N are fixed, it is known that the expected value $E(C)$ and the variance $V(C)$ of C are given respectively by

$$E(C) = Np_1p_2$$

and

$$V(C) \simeq Np_1q_1p_2q_2$$

where

$$p_1 = \frac{C + N_1}{N}; \qquad p_2 = \frac{C + N_2}{N} \quad \text{and} \quad p_1 + q_1 = p_2 + q_2 = 1.$$

Under the same conditions, the variance $V(\widehat{N})$ of $\widehat{N}$ is

$$V(\widehat{N}) = (C + N_1)^2(C + N_2)^2V\left(\frac{1}{C}\right)$$

which by the application of the formula given above reduces to

$$V(\widehat{N}) \simeq \frac{Nq_1q_2}{p_1p_2}\,.$$

The standard error of N is therefore

$$\sigma_{\widehat{N}} = \sqrt{\frac{Nq_1q_2}{p_1p_2}}$$

approximately.