Canadian Studies in Population, Vol. 7, 1980, pp. 1-7

A METHOD OF ESTIMATING THE TOTAL NUMBER OF EVENTS FROM INFORMATION SUPPLIED BY SEVERAL AGENCIES

Aleyamma George
Center for Mathematical Sciences, Trivandrum, Kerala, India
and
A. M. Mathai
McGill University, Montreal, Quebec, Canada

Résumé — Dans cette étude il s'agit d'une méthode d'estimation du nombre total d'événements en se fondant sur les données fournies par k agences différentes. La méthode est fondée sur le concept de divergence dirigée de la Théorie de l'Information. Une mesure convenablement choisie de divergence dirigée est réduite au minimum pour obtenir l'estimation. On a constaté que cette méthode-ci est plus facile à utiliser en comparaison avec les autres méthodes disponibles pour résoudre ce problème. On a expliqué la méthode en utilisant des données réelles.

Abstract — This paper deals with a method of estimating the total number of events based on the data supplied by k different agencies. The method is based on the concept of directed divergence in Information Theory. A suitably chosen measure of directed divergence is minimized to obtain the estimate. This method is seen to be easier to handle compared to the other methods available for this problem. The method is illustrated by using real life data.

Key Words — **estimation, new method, directed divergence, data from several agencies**

Introduction

Suppose k agencies collect data on certain events such as births or deaths or pregnancies, in a large segment of population, over a certain period. It happens very often that the agencies do not agree with each other's findings. For example, consider the enumeration of the total number of births in a certain region over a certain period. Suppose N is the true total number of births. We assume that the agencies are conducting the surveys independently and reporting them independently. Some of the total births may be correctly detected by the first agency alone, some by the second agency, some by the first and third agency alone and so on and some may be missed by all the agencies. We must estimate the true total number of births N based on the data supplied by all these k agencies.

The Method

For simplicity let us consider two agencies conducting the survey independently. Let p_1 and p_2 be the true probabilities of the first and second agencies respectively, detecting a birth. Then $q_1=1-p_1$, $q_2=1-p_2$ are the respective probabilities for not detecting the event — birth in our example. Let C_1 be the number of events detected by the first agency only. Under the assumption of independence the corresponding probability is

p_1q_2. Let C_2 be the number of events detected by the second agency alone, with probability q_1p_2. Let C_{12} be the number of events detected by both the agencies with probability p_1p_2. Then $N-C_1-C_2-C_{12}$ is the number of events missed by both the agencies with probability q_1q_2, where N is the true total number of events. Our aim is to estimate N. Even though C_1, C_2 and C_{12} are known, the number of events missed by both the agencies is unknown and hence N itself is unknown.

Chandrasekhar and Deming (1949) considered the problem of two agencies and proposed an estimate for N. Chakraborty (1963) extended the result to the case of three agencies and Das Gupta (1964) considered the general case of k agencies and obtained the estimate of N for the general case. He has shown that his estimate agrees with the estimate given by Chakraborty (1963) and Chandrasekhar and Deming (1949) for $k=2$. Both Chakraborty and Das Gupta used the method of maximum likelihood in arriving at the estimates after making approximations by neglecting higher powers of $1/N$. The estimate proposed by Chakraborty is:

$$\hat{N} = \left[\frac{n_1 n_2 \ldots n_k}{C_{12\ldots k}}\right]^{1/(k-1)} \tag{1}$$

where n_i is the total number of events recorded correctly by the ith agency, $i=1, \ldots, k$ and $C_{12\ldots k}$ is the number of events common to all the agencies. In our example of two agencies, $n_1=C_1+C_{12}$ and $n_2=C_2+C_{12}$. Das Gupta (1964) gave a polynomial equation in N for the general case and wrote down the explicit expressions for the cases $k=2$ and $k=3$. The procedure is simple. For example, for the case of $k=2$ we have four classes and from the multinomial distribution we have the likelihood function:

$$L = \frac{N!}{C_1!C_2!C_{12}!(N-C_1-C_2-C_{12})!}(p_1q_2)^{C_1}(p_2q_1)^{C_2}(p_1p_2)^{C_{12}}(q_1q_2)^{a} \tag{2}$$

where $a = N-C_1-C_2-C_{12}$

By maximizing L with respect to the unknowns p_1, p_2 and N the estimate of N is obtained. For convenience the exponents of p_1, q_1, p_2 and q_2 in L may be combined to obtain:

$$(p_1q_2)^{C_1}(p_2q_1)^{C_2}(p_1p_2)^{C_{12}}(q_1q_2)^{N-C_1-C_2-C_{12}}$$
$$= p_1^{n_1} p_2^{n_2} q_1^{N-n_1} q_2^{N-n_2} \tag{3}$$

where $n_1=C_1+C_{12}$ and $n_2=C_2+C_{12}$. Similarly when there are k agencies we have 2^k classes for the multinomial distribution.

For a fixed N we have a multinomial distribution and hence for a given N we have some justification in taking the estimates $\hat{p}_i=n_i/N$ and $\hat{q}_i=(N-n_i)/N$, $i=1, 2, \ldots, k$. But for a given p_i, $i=1, 2, \ldots, k$ we will look for alternate methods of estimating N.

The method proposed in this paper is based on the measure of directed divergence in Information Theory. For a detailed discussion of the various measures in Information Theory, their axiomatic definitions and applications, see Mathai and Rathie (1975). For the sake of completeness we will give a brief outline of the measure of directed divergence. Consider a discrete distribution $(p_1', p_2', \ldots, p_k')$, that is, consider a set of k mutually exclusive and totally exhaustive events $A_1, \ldots, A_k$ with the corresponding probabilities $p_1', p_2', \ldots, p_k'$ such that $p_i' \geq 0$, $i=1, \ldots, k$ and $p_1' + p_2' + \ldots + p_k' = 1$. Let these be the true probabilities of the events $A_1, \ldots, A_k$ and let $q_1', \ldots, q_k'$ be the probabilities assigned to $A_1, \ldots, A_k$ by an observer. As an example the $p_1', \ldots, p_k'$ may be unknown

and $q_1', \ldots, q_k'$ may be the relative frequencies taken as estimates for the unknown p_i', $i=1, \ldots, k$. Evidently $q_i' \geq 0$, $i=1, \ldots, k$ and $q_1' + \ldots + q_k' = 1$. The measure of directed divergence between the vectors $(p_1', \ldots, p_k')$ and $(q_1', \ldots, q_k')$ is:

$$D = \sum_{i=1}^{k} p_i' \log(p_i'/q_i') \tag{4}$$

with the usual convention that when a q_i' is zero the corresponding term is taken as zero. This D is a convenient measure to measure a type of divergence between $(p_1', \ldots, p_k')$ and $(q_1', \ldots, q_k')$.

The events are assumed to be mutually exclusive and totally exhaustive. In a practical situation the experimenter may be interested in a set of events which may not enable a partitioning of the sample space into mutually exclusive events. There may be other situations of statistical dependence where the probability of the intersection of two events of interest may not be equal to the product of individual probabilities. The following discussion does not deal with problems of various kinds of dependence. A detailed discussion of this aspect is undertaken by Marks, Seltzer and Krotki (1974).

Before introducing a criterion of estimation, a few words of caution may be in order. There are a number of standard methods of estimation available in the literature. Each method is motivated by different considerations. Practical situations often vary, hence not all methods are equally good or equally applicable in a given situation. All aspects of an experimental situation should be examined before selecting the most appropriate method. Usually, a practical situation is described by a set of basic assumptions. Investigation of the unique measures or methods which can result from a given set of assumptions can yield the best method of estimation. Such results are known as characterization theorems. An illustration of such results dealing with concepts in Information Theory and Statistics is available in Mathai and Rathie (1975). However, there is no method, which may be "good" in some situations, that is at the same time universally applicable.

In practical applications the experimenter is often tempted to use the popular methods such as the method of maximum likelihood or the method of moments. Such an approach can lead to misleading conclusions. For example, a given datum such as the waiting time for the first conception in a certain group of females may appear to fit a gamma distribution. One may estimate the two parameters in the gamma density by using the method of moments or the method of maximum likelihood. Each method gives a pair of estimates for the parameters, which are fitted to a gamma distribution and tested for goodness of fit. One may reject the hypothesis of "good fit" for all estimates of the parameters given by the methods available to him. If one concludes that a gamma distribution is not a good fit to the data his conclusions may be invalid. There may be a member in the same gamma family which is not obtained by any standard method of estimation. Illustration of this fact, using real life data is given in George and Mathai (1976).

From the above discussions it is evident that when a method of estimation is proposed, one can point out the basic assumptions and the motivating factors, but the most important thing to remember is that additional conditions or assumptions should not be made during mathematical manipulations which might make that estimation method meaningless, or which might contradict the basic assumptions of the practical situation. A defect of this type will be illustrated later in the discussion, where the method of maximum likelihood is used in estimating certain events.

The method of estimation proposed in this article is based on the measure of directed

divergence defined in (4). When estimating a hypothetical vector by using an observed vector our aim is to minimize the distance between the two. A measure of directed divergence is a type of distance between two discrete distributions satisfying most of the postulates for a mathematical distance. Hence a method of estimation based on directed divergence is well motivated. Further, this measure is widely applied in a range of areas such as communication theory, and psychology. This enlarges the scope of borrowing of useful results and techniques from these various disciplines for application to population problems.

Consider the case of two agencies. We have four mutually exclusive and totally exhaustive events with probabilities p_1q_2, p_2q_1, p_1p_2 and q_1q_2. Let these be denoted by $p_1' = p_1q_2$, $p_2' = p_2q_1$, $p_3' = p_1p_2$, and $p_4' = q_1q_2$. Evidently $p_i' \geq 0$, $i=1, 2, 3, 4$ and $p_1' + \ldots + p_4' = 1$. Naturally an observer will assign relative frequencies C_1/N, C_2/N, C_{12}/N, $(N-C_1-C_2-C_{12})/N$ to p_1', p_2', p_3' and p_4' respectively. Hence the directed divergence:

$$\begin{aligned} D &= p_1' \log\{p_1'(N/C_1)\} + p_2' \log\{p_2'(N/C_2)\} + p_3' \log\{p_3'(N/C_{12})\} \\ &\quad + p_4' \log\{p_4'(N/(N-C_1-C_2-C_{12}))\} \\ &= (p_1q_2) \log\{(p_1q_2)(N/C_1)\} + (p_2q_1) \log\{(p_2q_1)(N/C_2)\} \\ &\quad + (p_1p_2) \log\{(p_1p_2)(N/C_{12})\} \\ &\quad + (q_1q_2) \log\{(q_1q_2)(N/(N-C_1-C_2-C_{12}))\} \\ &= \log N - (q_1q_2) \log(N-C_1-C_2-C_{12}) + (p_1q_2) \log((p_1q_2)/C_1) \\ &\quad + (p_2q_1) \log((p_2q_1)/C_2) + (p_1p_2) \log((p_1p_2)/C_{12}) \\ &\quad + (q_1q_2) \log(q_1q_2) \end{aligned} \tag{5}$$

Only two terms contain N in (5) and minimizing D with the help of calculus we have the turning value given by

$$(1/N) - (q_1q_2)/(N-C_1-C_2-C_{12}) = 0 \tag{6}$$

That is:

$$\hat{N} = (C_1+C_2+C_{12})/(1-q_1q_2) \tag{7}$$

Thus for given values of q_1 and q_2 the minimum directed divergence estimate of N is $(C_1+C_2+C_{12})/(1-q_1q_2)$. In the case of $k=3$ the terms containing N are $\log N - (q_1q_2q_3) \log(N-C_1-C_2-C_3-C_{12}-C_{13}-C_{23}-C_{123})$. Hence proceeding in the same way as before we have

$$\hat{N} = (C_1+C_2+C_{12}+C_{13}+C_{23}+C_{123})/(1-q_1q_2q_3) \tag{8}$$

with the corresponding notations. Similarly in the general case of k agencies the estimate will be

$$\hat{N} = (C_1+C_2+\ldots+C_k+C_{12}+\ldots+C_{12\ldots k})/(1-q_1q_2\ldots q_k) \tag{9}$$

But when N is given we have seen that the estimates of q_i's are available from the multinomial law, that is,

$$\hat{q}_i = (N-n_i)/N, \; i=1, \ldots, k \tag{10}$$

where n_i is the total number of events recorded correctly by the ith agency. Thus we propose the following estimate

$$\hat{N} = (C_1+C_2+\ldots+C_k+C_{12}+\ldots+C_{12\ldots k})/ \left[1-\left(\frac{\hat{N}-n_1}{\hat{N}}\right)\left(\frac{\hat{N}-n_2}{\hat{N}}\right)\ldots\left(\frac{\hat{N}-n_k}{\hat{N}}\right)\right] \tag{11}$$

in implicit form which, when solved, is the estimate. It should be pointed out here that we did not use any large N approximations to arrive at the above estimate.

Particular cases

When $k=2$, the expression in (11) reduces to

$$\hat{N} = (C_1+C_2+C_{12})/[1-(\frac{\hat{N}-n_1}{\hat{N}})(\frac{\hat{N}-n_2}{\hat{N}})] \tag{12}$$

That is,

$$1 = \hat{N}(C_1+C_2+C_{12})/[(n_1+n_2)\hat{N} - n_1n_2]$$
$$\hat{N} = n_1n_2/C_{12} \tag{13}$$

which agrees with the estimate proposed by Chandrasekhar and Deming (1949), Chakraborty (1963) and Das Gupta (1964). When $k=3$, the estimate of N in (11) reduces to

$$\hat{N} = (C_1+C_2+C_3+C_{12}+C_{13}+C_{23}+C_{123})/[1-(\frac{\hat{N}-n_1}{\hat{N}})(\frac{\hat{N}-n_2}{\hat{N}})(\frac{\hat{N}-n_3}{\hat{N}})] \tag{14}$$

where as before $n_1=C_1+C_{12}+C_{13}+C_{123}$, $n_2=C_2+C_{12}+C_{23}+C_{123}$, and $n_3=C_3+C_{13}+C_{23}+C_{123}$. Solving (14) for N, we get, after simplification,

$$\hat{N} = [b+(b^2-4ac)^{1/2}]/(2a) \tag{15}$$

where

$$a = C_{12}+C_{13}+C_{23}+2C_{123}$$
$$b = n_1n_2 + n_1n_3 + n_2n_3$$
$$c = n_1n_2n_3$$

The estimate of N in (15) agrees with that obtained by Das Gupta (1964).

A Comparison

The estimate of N for the general case, in implicit form, is given in (11) which when simplified reduces to the form

$$1 = C(\hat{N})^{k-1}/[\hat{N}^k - \prod_{i=1}^{k}(\hat{N}-n_i)] \tag{16}$$

or,

$$\prod_{i=1}^{k}(\hat{N}-n_i) - (\hat{N}-C)(\hat{N})^{k-1} = 0$$

where

$$C = C_1+C_2+\ldots+C_k+C_{12}+\ldots+C_{12\ldots k}$$

This is exactly the same as equation (2.6) of Das Gupta (1964). He, however, obtained his equation after omitting terms such as $1/N$, $1/N^2$, . . ., and $1/(N-C)$, $1/(N-C)^2$, ... It should be remarked that it is least desirable to omit terms such as $1/(N-C)$, $1/(N-C)^2$, . . ., from a practical point of view. If $N-C$ is assumed to be large, by implication, the number of events missed by all the agencies is also large, which makes the procedure of estimation under this assumption questionable. If the method of maximum likelihood is used, then a reasonable and valid procedure is to take all the terms and solve the normal equation to obtain the turning points. The choice of a particular method should be decided by other considerations, but once chosen, conditions should not be added, such as the one above, which could nullify or overshadow the problem itself. From the nature of the expression involved, one can see that the procedure of maximum likelihood does not yield to manageable equations and the approximations are not justifiable in the problem under consideration. The method proposed in the present paper eliminates such difficulties.

A Real-Life example

As an illustration, Das Gupta (1964) used simulated data from a model sampling experiment. Here we will use a real life example. The following data relate to the birth register kept by the Registrar-General of India, Supervisor of Records, and data collected by independent investigators under a PL-480 scheme, University of Kerala. For convenience, the data are given in Table 1.

TABLE 1 BIRTHS RECORDED FOR THE PERIOD JANUARY 1, 1966 TO JUNE 30, 1966 BY DIFFERENT AGENCIES

Sample area	C_1	C_2	C_3	C_{12}	C_{13}	C_{23}	C_{123}	Total
Karuvilakom	-	-	4	2	-	5	16	27
Nalloorvattom	-	-	1	3	2	7	8	21
Venkatambu	-	-	1	5	-	-	14	20
Kuzhakkad	-	1	6	-	-	4	11	22
Kadayara	-	1	-	6	-	-	33	40
Madathuvilakom	-	1	6	1	-	4	6	18
Kizhakumkara	-	-	6	1	-	1	28	36
Pallikkal	-	-	2	-	4	-	24	30
Ottasekhara-mangalam	-	-	1	-	-	9	7	17
Total	-	3	27	18	6	30	147	231

Total number of births reported by the registrar:	$(C_1+C_{12}+C_{13}+C_{123})$	= 171
Total number of births reported by the supervisor:	$(C_2+C_{12}+C_{23}+C_{123})$	= 198
Total number of births reported by the investigators:	$(C_3+C_{13}+C_{23}+C_{123})$	= 210
Total number of births missed by the registrar:	$(C_2+C_3+C_{23})$	= 60
Total number of births missed by the supervisor:	$(C_1+C_3+C_{13})$	= 33
Total number of births missed by the investigators:	$(C_1+C_2+C_{12})$	= 21

In Table 2 we present the estimated values of N based on Chandrasekhar and Deming formula taking two by two, Chakraborty's formula for $k=3$, Das Gupta's and our formula for $k=3$.

It may be seen from Table 2 that the invesigators have recorded 12 more births in total than the supervisor. When the supervisor has recorded 27 more births, the investigators have recorded 39 more births than the registrar.

A comparison of the estimates shows that the Chandrasekhar-Deming estimates based on the records of the registrar and investigators, and Das Gupta's estimates and George & Mathai's estimates based on all the three agencies are closer to each other.

TABLE 2 ACTUAL NUMBER OF BIRTHS ENUMERATED AND ESTIMATED VALUES FOR THE PERIOD FROM JANUARY 1, 1966 TO JUNE 30, 1966

Sample area	Actual number enumerated by			Chandrasekhar-Deming estimates based on		(1)	(2) & (3)
	(a)	(b)	(c)	(a) & (b)	(a) & (c)		
Karuvilakom	18	23	25	23	28 (12.02)	25	27 (11.59)
Nalloorvattom	13	18	18	21	23 (9.87)	23	21 (9.01)
Venkatambu	19	19	15	19	20 (8.58)	20	20 (8.58)
Kuzhakkad	11	16	21	16	21 (9.01)	18	22 (9.44)
Kadayara	39	40	33	40	39 (16.74)	40	40 (17.17)
Madathuvilakom	7	12	16	12	19 (8.15)	15	19 (8.15)
Kizhakkumkara	29	30	35	30	36 (15.45)	33	36 (15.45)
Pallikkal	28	24	30	28	30 (12.88)	29	30 (12.88)
Ottasekhara-mangalam	7	16	17	16	17 (7.30)	17	18 (7.73)
Total	171	198	210	205	233 (100)	220	233 (100)

Figures in brackets are percentages: (a)=Registrar, (b)=Supervisor, (c)=Investigators; (1)=Chakraborty's estimate, (2)=Das Gupta's estimate, (3)=George and Mathai's estimate.

References

Chakraborty, P.N. 1963. On a method of estimating birth and death rates from several agencies. Calcutta Statistical Association Bulletin 12:106-112.

Chandrasekhar, C. and W. E. Deming. 1949. On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association 44:101-115.

Das Gupta, P. 1964. On the estimation of the total number of events and of probabilities of detecting an event from information supplied by several agencies. Calcutta Statistical Association Bulletin 12:89-100.

George, Aleyamma. 1968. Survey Methods of estimating vital events. (Survey report).

George, A. and A. M. Mathai. 1976. Distribution of birth intervals based on original data. Demography of India 5:163-180.

Marks, E. S., W. Seltzer, and K. J. Krotki. 1974. Population Growth Estimation: A Handbook of Vital Statistics Measurements. The Population Council.

Mathai, A. M. and P. N. Rathie. 1975. Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications. New York and New Delhi: Wiley.

Received December, 1978; revised November, 1979.