



## CME

### Estimating Hidden Events

Suprakas Hazra\*

In order to estimate any event either on health care e.g. births, deaths, morbidity, disability or any events on historical, social data etc. through any study by one set of source, the questions which naturally arises in mind as to what proportion of the events are missed. What is the extent of under registration? Does it really need to have other independent source to verify it before taking any action? Is it really possible to estimate the accurate measure of the event in which we are interested and if so at what level of accuracy?

Patama Vapattanawong (2011)<sup>1</sup> & Pramote Prasartkul, in their study showed that in Thailand, overall under registration of deaths during 2005-06 was 8.69% for males and 8.36% for females. For both males & females, under registration decreased as age increased. Under registration was greater among people of either sex aged 1-4 years, whereas it was less than 10% among people 60 years of age and older, both males and females. Under registration was estimated by cross matching two lists of deaths obtained from two independent agencies. The matching criteria were on the basis of name, sex, age of diseased, place of death, place of residence etc.

Crimmins, E (1980)<sup>2</sup>, in her study had shown that death registration and mortality enumeration in the Census of 1900 in U.S, rural counties that under enumerated deaths also intended to under register them (In U.S. the mortality data for almost all states from 1850 through 1900 as part of the Federal decennial census were not used because of unreliability and deficiency. Crimmins had used the data and provided estimates of the completeness of the available mortality data for 1900 collected by the two processes of registration and enumeration)

---

**\* Faculty, Department of Epidemiology, All India Institute of Hygiene & Public Health, Kolkata.**

Douglas 1999<sup>3</sup> in a study showed dramatic result in estimating the number of unrecorded murders, using newspaper and archives data for the state of South Carolina, 1877-1878 using multisource 'Capture – recapture' (or 'dual enumeration') method. He had shown in comparison with the capture – recapture estimate of total homicide, at least 58% of the State's murders for the 2 years were not found in the South Carolina, State Department of Archives and History, the major newspaper of the State missed at least 30% and the Combined sources missed at least 20%.

Therefore it is observed that to estimate the under registration of the vital events or to provide the completeness of the mortality data or to estimate the unrecorded murders ,two independent sources are required. Hence estimating under registration of any event is a major part to complete a study.

In all the situations above, the extent of the hidden events was estimated using a technique and formula which was developed by Professor

#### **Chidambara Chandrasekaran and William Edwards Deming.**

**William Edwards Deming** (October 14, 1900 – December 20, 1993) was an American engineer, statistician, professor, author, lecturer, and management consultant. Educated initially as an electrical engineer and later specializing in mathematical physics, he helped develop the sampling techniques still used by the U.S. Department of the Census and the Bureau of Labor Statistics..... )

(From Wikipedia )

**Chidambara Chandrasekaran**, (1911–2000), was noted Indian demographer and statistician, was educated in India, UK and the US. He graduated from Morris College, Nagpur, with a B.Sc. degree, followed by a M.Sc. degree from the Nagpur University, and a PhD degree in Statistics from University College London in 1938. He was also awarded an MPH degree from Johns Hopkins School of Hygiene and Public Health in 1947. He was related to two Nobel Prize winners: C. V. Raman was his uncle and Subrahmanyan Chandrasekhar was his cousin.

He worked for the United Nations and World Bank in various capacities. He was elected as the President of International Union for the Scientific Study of Population (IUSSP) and served in this position from 1969-73. He held academic position at various Indian universities. He was a Professor of Biostatistics at the **All India Institute of Hygiene and Public Health, Calcutta**, from 1941-48 and 1954-58. This was followed by a stint as the Director of the Demographic Training and Research Centre, Mumbai (later renamed as the International Institute of Population Sciences) from 1959-64.

One of his most important contributions to the field of demography was in developing a technique to estimate the number of vital events by comparing results from two different systems (such as a sample survey and a vital registration system). This technique is commonly known as Chandra-Deming formula (also known as Dual Record System) and was first proposed in an article in 1949<sup>4</sup> "On a method of estimating birth and death rates and the extent of registration." *Journal of the American Statistical Association*, 44(245): 101-115 (co-authored with W. Edwards Deming). Various improvements and adaptations of this method are now commonly used in developing countries, including India, to estimate birth and death rates.

Chandrasekaran was the lead investigator of the Mysore Population Study, funded by the United Nations and the Indian Government, which was a pioneering survey in collecting fertility related information, including contraceptive use, in a developing country and in demonstrating that such data could be used for analyses of fertility determinants. He also investigated the population change of the Parsis in India and the reproduction patterns of Bengali women.

Chandrasekaran was active promoter of family planning policies in India and at least on one occasion advised Jawaharlal Nehru, former Prime minister of India, on matters of demographic transition.

(From Wikipedia)

Chanrasekharan and Deming (1949) developed the method to estimate the number of births or deaths missed by the Registrars . To estimate the missing number of events they matched individuals from two independent lists originally a Registrar's list and the list obtained via

house to house canvass, called “Investigator’s list ( Dual enumeration ) Each list is supposed to miss some births or deaths and each serves as a criteria for judging the completeness of the other . The basic characteristics of this method allowed estimation of the “dark figure “ and in some cases standard errors and confidence interval as well.

In the development of the theory, allowance was made for the fact that the chance of an event being missed by one list (” Registrar ’s list or 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. This is done by sub dividing 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 number of events in these groups separately and summing them for a total.

The theory was applied to an enquiry covering the year 1945 and 1946 separately in February 1947 over Singur service area under Rural Health Unit & Training Centre, Hooghly, a rural unit of All India Institute of Hygiene & Public Health, Kolkata. It was found that the estimated total number of events for the area was greater than the number of events collected by the two independent agencies.

**Derivation of formula –**

Let

R: The list prepared by the Routine Registrar

I : The list prepared by the Investigator

C: Number of events registered by both” R” &” I “

N<sub>1</sub>: Number of events registered ” R” only ( Not in “I”)

N<sub>2</sub>: Number of events registered by ” I” only ( Not in “R”)

Y: Number of events missed by both” R” &” I “

N: number of events actually occurred

The objective is to estimate Y and subsequently N

Based on the above information following table can be constructed

List prepared by	Number of events detected	Number of events missed
R	C+N <sub>1</sub>	N-(C+N <sub>1</sub> )
I	C+N <sub>2</sub>	N-(C+N <sub>2</sub> )

Probability that an event is detected by 'R' ( p<sub>1</sub> ) = (C+ N<sub>1</sub>) / N

Probability that an event is detected by " I " ( p<sub>2</sub> ) = (C+ N<sub>2</sub>) / N

Probability that an event is missed by " R" ( q<sub>1</sub> ) = 1- (C+ N<sub>1</sub>) / N = (N-C- N<sub>1</sub>)/ N

Probability that an event is missed by " I" ( q<sub>2</sub> ) = 1- (C+ N<sub>2</sub>) / N = (N-C- N<sub>2</sub>)/ N

Since the events are collected by two independent agencies, therefore the probability that an event is missed by both" R" &" I " is

$$q_1 q_2 = \frac{(N-C-N_1)}{N} \cdot \frac{(N-C-N_2)}{N}$$

Therefore expected number of events missed by both" R" &" I "

$$\begin{aligned} & \frac{N(N-C-N_1)(N-C-N_2)}{N \cdot N} \\ &= \frac{(N-C-N_1)(N-C-N_2)}{N} \end{aligned} \quad \text{----- (1)}$$

Again the number of distinct events collected by both" R" &" I " = C+ N<sub>1</sub> + N<sub>2</sub>

Y = The number of distinct events missed by both" R" &" I " = N- (C+ N<sub>1</sub> + N<sub>2</sub>) ..... ( 2)

Equating ( 1) and ( 2)

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

or , (N-C- N<sub>1</sub>) (N-C- N<sub>2</sub>) = N(N-C- N<sub>1</sub> - N<sub>2</sub>)

or,  $N_1 N_2 = C(N - C - N_1 - N_2)$

or,  $N_1 N_2 = C Y$

or,  $Y = N_1 N_2 / C$

Hence the estimated number of events missed by both "R" & "I"  $(Y) = N_1 N_2 / C$   
and subsequently estimated number of events  $(N) = C + N_1 + N_2 + N_1 N_2 / C$

It can be shown that estimated value of N is the unbiased estimate in the limit when N becomes large and the assumption just mentioned is valid.

The standard error of  $N = \sqrt{N q_1 q_2 / p_1 p_2}$

95% Confidence interval of  $N = N \pm 1.96 \sqrt{N q_1 q_2 / p_1 p_2}$

Unfortunately such an important and universally used technique and formula is not being used much in India. Scott C<sup>5</sup> mentioned "... This method of estimating the missing element is known as the "CSD estimation" after Chandrasekhar and Deming who developed it. Only India does not follow this practice of estimating missing events.....'.

One of the reasons for non-application of this technique and formula may be that this is neither available in any book of Statistics nor in the website too. Presently no one has the access to this article rather than purchasing with USD -186 for the issue or USD -46 for the article. Author affiliations –AIHH& PH ,Kolkata .The objective of the author is to highlight the technique and formula of Chandrasekhar and Deming with a high anticipation that it will be used by the researchers not only to estimate the vital events but to estimate other hidden cases also in various fields of study.

### References:

1. PatamaVapattanawong ( 2011) &PramotePrasartkul- Under Registration of deaths in Thailand in 2005 – 2006, results of cross-matching data from two sources ,  
Publication : Bulletin of World Health Organisation ; Type ; Research article ID  
:BLT.10-08.083931

2. Crimmins E. (1980) The completeness of 1900 mortality data collected by registration and enumeration for rural and urban parts of states: Estimates using Chandrasekar – deming technique. *Historical methods*. 13.163-169
3. Douglas L. Eckberg (1999) . A capture – recapture approach to estimation of hidden historical killings. *The varieties of homicides and its research* , Chapter- Methodology of Historical studies , chapter one , 1-10
4. Chandrasekhar C. Deming, W.E. (1949) . On a method of estimating birth and death rates and the extent registration. *journal of the American Statistical association* , 44, 101-11
5. Scott C, In: *International Union for the Scientific Study of Population. International Population Conference, Liege, 1973. Vol. 2. Liege, Belgium, IUSSP, 1973. 407-16. POPLINE HEALTH. The dual record (PGE) system for vital rate measurement: some suggestions for further development*