Estimation of COVID-19 Prevalence under Closed Population using Capture-Recapture approach

Rakesh Kumar[1], R.S.Mishra[2]

Delhi Technological University, Delhi, India

Abstract: *The nation-wide lockdown due to COVID-19 created the situation of closed population. This paper presents a general theory for estimating the size of a closed population. The theory is based on a log-linear model, with model parameters representing the natural logarithm of subpopulation size and the natural logarithm of subpopulation capture probabilities. Lastly the paper contains illustration of use of log-linear to closed population.*

Keywords: COVID-19, log-linear, capture-recapture

1. Introduction

Estimating population sizes has long been of interest, from the estimation of the human or ecological population size within regions or countries to the hidden number of corona patients in the state. Total enumeration of the population, for example, via a census, is often infeasible or simply impractical. However, a series of partial enumerations or observations of the population is often possible. This has led to the ideas of capture-recapture methods, which have been extensively used within ecology to estimate the size of wildlife populations, with an associated measure of uncertainty, and are most effectively applied when there are multiple capture occasions. The capture-recapture technique originally developed for animal studies and later has been applied by practitioners and researchers to human populations under the term `multiple-record system' for an arbitrary number of lists and under the term `dual-system especially for two lists. The earliest references to such applications can be traced back to Sekar & Deming (1949) for two samples, Wittes & Sidel (1968) for three samples, Wittes (1974) for four samples and Fienberg (1972) for five samples.

Since state circumstances will vary in terms of disease activity and capacity to respond, caution should be exercised in directly linking severity assessment at a global to national level to actions at the state level. It is likely that information will be limited early in the pandemic while the demand for information simultaneously escalates. If pandemic surveillance is to provide sufficient information and data to assess severity, country/state need to review their existing surveillance procedure and capacity to address the weaknesses to be prepared for pandemic surveillance. Essential components of an effective pandemic corona surveillance system will include: early detection and investigation; comprehensive assessment; and Monitoring. Delhi government launched 5T plan to stop the spread of the novel corona virus. The 5T plan includes Testing, Tracing, Treatment, Teamwork and Tracking based on the South Korean strategy Test, Trace and Isolate. But the step testing requires basis. India's apex medical body Indian Council of Medical Research (ICMR) intensified the random sampling of people who display flu-like symptoms but don't have any history of travel to outbreak zones to determine whether community transmission is taking place. Later ICMR experts

aren't convinced about following South Korea's strategy, which has reported more recoveries from Covid-19 than new infections through aggressive testing without lockdowns. Indian experts said they will revise the testing protocol depending on the situation. The ICMR in its study highlighted that Covid-19 containment activities need to be targeted in districts reporting positive cases among Severe Acute Respiratory Infections (SARI) patients and stated that intensifying sentinel surveillance for Covid-19 among SARI patients may be an efficient tool to effectively use resources towards containment and mitigation. Testing: Until now testing was very limited. Only symptomatic people who had history of travel were tested. But now Delhi government plans to aggressively start, testing drives in Delhi areas which are being seen as Corona virus hot spot. Also there is going to be random testing of one lakh people in the next few days. Delhi is the first state to carry out random testing. Second T - Tracing as testing ramps up every single contact of positive cases through cellphone triangulation will be traced. Twenty four hour tabs will be kept on those who are in quarantine to ensure it's being followed. Such focused actions will help in staying at least two steps ahead of virus. ICMR (Indian Council of Medical Research) conducted a test in few states of country and its research shown that there are many cases of corona virus where it is difficult to find the source i.e. they have no travel history and not get in contact with any such person who has travel history. The Ministry of Home Affairs (MHA) of India stated that the government has adopted a special strategy to fight coronavirus, with specific instructions for the district and state authorities. The strategic approach for COVID-19 containment broadly involves defining the area of operation, applying perimeter control, delineating containment and buffer zones. Besides, there is active search for cases, contact tracing, quarantine, clinical management and awareness generation among public.

Without discussing the methodology, It is estimated by Indian professor Ramanan Laxminarayan, Director and Senior Fellow, The Center for Disease Dynamics, Economics & Policy that by July 2020, 300 to 500 million Indians will get infected and a couple of million Indians can die of corona virus cases and fatalities. A data journalist Rukmini S. Has also presented her observation on projections of corona virus cases concluded that if the rate of COVID-19 spread continues, 30,000 deaths can occur in

India by May 2020 and also mentioned that various variables like lockdowns in India and the enforcement of them are great challenges that statisticians cannot deal with.

Random Sampling

Random sampling is basic to many computing applications in science, statistics, and engineering. Random sampling is used on those occasions when processing the entire dataset is unnecessary and too expensive in terms of response time or resource usage. The savings generated by sampling may arise either from reductions in the cost of retrieving the data from the database or from subsequent "post-processing" of the sample. Sampling is useful for applications which are attempting to estimate some aggregate property of a set of records, such as the total number of records which satisfy some predicate. Thus random sampling is typically used to support statistical analysis of a dataset, either to estimate parameters of interest or for hypothesis testing. Random sampling has three common justifications: reduced database retrieval costs (or time), reduced "post-processing" computational costs, or reduced cost of physical inspections. So, random sampling is a basic sampling technique through which each individual is chosen entirely by chance and each member of the population has an equal probability of being included in the sample. It is the only way to be sure that the sample is not biased and is representative of the population as a whole, and even more complex sampling schemes such as systematic and stratified sampling are based on this basic scheme. ICMR's testing of a small random sample may not be able to capture it. As India is divided its 755 districts into green, orange and red zones. It will entail making sure that green zones are indeed green by deploying testing strategies. Delhi government announced a five-point action plan of "five Ts -- testing, tracing, treatment, team-work and tracking -- and monitoring" to contain the virus spread and said one lakh random rapid anti-body blood tests for COVID-19 tests will be conducted in the city's hotspot areas, the similar plan is followed by other states also.

2. Literature Review

The past literature is reviewed keeping in view for this study is concerned as well as COVID-19. Pandemic is not new for the world, but the lockdown concept is new for country like India experience in year 2020. Estimating population sizes has long been of interest, from the estimation of the human or ecological population size within regions or countries to the hidden number of corona patients in the state. Total enumeration of the population, for example, via a census, is often infeasible or simply impractical. However, a series of partial enumerations or observations of the population is often possible. This has led to the ideas of capture-recapture methods, which have been extensively used within ecology to estimate the size of wildlife populations, with an associated measure of uncertainty, and are most effectively applied when there are multiple capture occasions.

Capture-recapture methods are used in various areas, from epidemiology (Coeli et al., 2000) to zoological studies of populations (Fernandez, 1995) or even in the analysis of the variation of measurement intervals of closed communities of diverse magnitudes (Cormack, 1992). The capture-recapture techniques are used for closed populations, where the effects of birth rate or mortality and migration are not taken into account, in this case it is considered no change in population size. They can also be used for open populations, where population changes caused by birth rate, mortality and migration are taken into account (Cormack, 1989). Sekar and Deming (1949) proposed their research concerned with estimating birth and death rates in an area near Calcutta, India. The research based report of Robles, S. C. et al. (1988) described about usefulness of capture-recapture methods used to estimate completeness at the Ontario Cancer Registry. The method was applied in two fashions; first, using three data sources in a modeling approach and second, using two data sources and standard, simple capture-recapture methods. The modeling approach is more flexible, since several variables that influence cancer registration can be considered and can be used to identify reporting patterns of different data sources. Wittes, J. T (1974) in his paper considered the problem of estimating the total size of a target population from which a study sample has been obtained by merging names from several routinely collected lists. Each list is assumed to be a simple random sample from the target population and all lists are assumed independent. Hook, E. B. (1980) presented his work with goal to compare various methods for estimating the live-birth prevalence of spina bifida (cystica) using Bernoulli census analysis, allowing for under ascertainment in the sources used, and to determine if a stable estimate or at least a useful confidence boundary on the population live-birth prevalence rate could be derived from prenatal diagnosis of spina bifida undertaken as part of a cost-benefit analysis. Brenner, H. (1994) highlighted the drawback of routine systems of disease monitoring, such as morbidity registries, are often prone to specific threats of validity, such as in imperfect diagnoses or imperfect record linkage. Through quantitative assessment of the performance of the two-source capture-recapture method for disease monitoring in the presence of imperfect record linkage. International Working Group for Disease Monitoring and Forecasting (1995) evaluates the potential epidemiologic uses of capture-recapture, which include the primary area of determining disease frequency. Capture-recapture may be a means to effectively "count" new cases (incidence) or count existing cases (prevalence). In addition to generating an estimate of population size, another application of capture-recapture is to assess the costs of ascertainment relative to the degree of accuracy. Fisher N. et al. (1994) worked on routinely collected data from local agencies to provide a reliable estimate of the size and psychiatric morbidity of the homeless population of a given central London by using capture-recapture analysis. A multiple sample, log-linear capture-recapture method was applied to a defined geographical area during 6 months. The method calculates the total homeless population from the sum of the population actually observed and an estimate of the unobserved population. Sheila M. Bird and Ruth King (2018) with changed name of Capture-Recapture Estimation with Multiple Systems Estimation (MSE) advocated that the MSE approach has been preferred when estimating "capture-shy" or hard to-reach populations, including those caught up in the criminal justice system; or homeless; or trafficked; or civilian casualties of war. The paper focuses on whether the uncertainty which qualified their estimates was sufficiently narrow to orient public policy or not, if not, what options

were available and/or taken to reduce the uncertainty or to seek external validation. *Jones H.E.et al.(2016) by using* Capture–recapture analysis recommended for estimating the prevalence of problem drug use or people who inject drugs .The work aim to demonstrate how naive application of Capture–recapture can lead to highly misleading results, and to suggest how the problems might be overcome. The method further validated by a case study of estimating the prevalence of people who inject drugs in Bristol, UK, using Capture–recapture. *Leclerc, P. (2014) already published similar type of work as Jones H.E.et al.(2016) carried but study area was* Island of Montréal, Canada, Using a Six-Source Capture-Recapture Model.

Even after worthiness of this method Stephen N Jarvis et al. (2000) stated with prior proof in his research paper with illustration that there is also serious drawbacks in the use of the mark/recapture technique when applied to injuries. Through example of two sample mark/recapture estimation in an epidemiological setting it is shown to violate virtually all the requirements for valid use of the technique. Further mentioned that very little can be deduced accurately about the scale or characteristics of an unobserved group by the use of mark/recapture applied to two overlapping health event registers. Begon, M. (1983) as a critic to the methods of Jolly he mentioned that there has also been a consistent interest in the method on the part of statisticians, a sporadic interest in field-based critiques of the method, and a considerable tendency, especially since the mid-seventies, to refer to Jolly's paper but make no use of its contents.

3. Capture-Recapture Method and Log-Linear Model

Pierre Simon Laplace (1749-1827) used similar to capture-recapture approach to estimate the size of human population of France. Further Lincoln-Peterson develop a technique with the name capture-recapture

Closed Population: A closed population is one in which the total number of individuals is not changing through births, deaths, immigration or emigration. But in practice, most real populations are not closed, it's a chance the in complete lockdown, the population is definitely closed. An open population is one that is (or could be) changing during the course of study, because of any combination of births, deaths, immigration or emigration. The 'closed captures' data types all assume the population of interest is closed during the sampling period (White *et al.* 1982). Meaning, the models assume that no births or deaths occur and no immigration or emigration occurs. Typically, we refer to a closed population as one that is free of unknown changes in abundance, as we can usually account for known changes.

The assumption is in context of human being is differ from animals in that way that- human being are self conscious and able to report about their illness and checked (tagged), human being always available whenever called for recounting. But this COVID-19 effected the mind and thought of people in such a way that people use to hide themselves from the society and that become a challenging task to the health department to do mapping to identify them from the huge crowd of population, even if it is closed due

to lockdown. The 'closed captures' data types all assume the population of interest is closed during the sampling period (White *et al.* 1982). Meaning, the models assume that no births or deaths occur and no immigration or emigration occurs. Typically, we refer to a closed population as one that is free of unknown changes in abundance, as we can usually account for known changes.
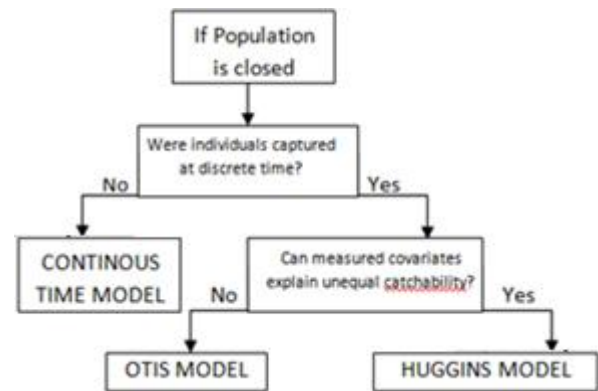

Figure 1: Flow Chart of the methods of CAPTURE-RECAPTURE

Consider a closed population with size N and the people are indexed by 1,2,. . . , N in any arbitrary fashion. In a usual t-sample capture-recapture experiment, people are captured and marked in each trapping sample (or trapping occasion), and we only notice in each sample whether each person was captured or not without recording the sequential order or the time when it was captured. Thus the capture history can be expressed as a N x t matrix (X_{ij}), where $X_{ij} = I$ [the ith person was caught in the jth sample], i = 1, 2,. . . , N, j = 1, 2.. . , t. Extensive models have been discussed for the above discrete-time type data.

If the capture time and the sequential order for each individual can be recorded, then a continuous-time model is needed. Anderson and Wilson (1989) have done some simulation to compare the existing estimators for continuous-time models. Becker (1984), Becker and Heyde (1990) proposed an estimation procedure using a martingale approach.

In the complete lockdown situation the whole population is considered as closed population and the new birth in this period is neglected because the record is unavailable for this period. A closed population is one in which the total number of individuals is not changing through births, deaths, immigrations and emigrations and all these were not in assumption in corona period. In practice, most real populations are not closed. Sometimes, however, the changes over the time period of interest are small enough that the assumption of closure is a reasonable approximation, and the effects of violating that assumption are minimal. For that reason, the analysis of capture-recapture data from closed populations continues to be a topic of interest to researchers.

However closure is subdivided into two components: (i) geographic and (ii) demographic. The geographic closure of population by a boundary, which limits the population, and demographic is about birth, death, immigration and emigration.

The models of Otis et al.(1978) denoted M_0, M_t, M_b, M_h, M_{bh}, M_{tb}, M_{th}, and M_{tbh}.

Model M_0 assumes a single common probability of capture. The subscripts t, b, and h correspond to temporal variation in common probability of capture, behavioral response to common probability of capture, and individual response to capture, respectively. Multiple subscripts indicate a combination of capture responses.

The above mentioned eight models summarised by symbol, source of variation in capture probabilities , and the associated estimators, if any , program CAPTURE uses theses names for estimators.

Table 1: Discrete-Time Models and Associated Appropriate Estimator with Log Linear form

Model	Source of variation in capture probabilities	Appropriate estimator	Log Linear form
M_0	None	Null (All estimators given below are valid)	$log\ (P_{ij}) = \alpha$
M_t	Time	Darroch (1958)	$log\ (P_{ij}) = \mu_j$
M_b	Behaviour	Zippin (1956)	$log\ (P_{ij}) = \alpha + ^{\gamma} Y_{ij}\ (\alpha_i = \alpha)$
M_h	Heterogeneity	Jackknife (1978)	$log\ (P_{ij}) = \alpha_i$
M_{bh}	behaviour heterogeneity	Generalized Removal (Otis et al. 1978)	$log\ (P_{ij}) = \alpha_i + ^{\gamma} Y_{ij}$
M_{tb}	time behaviour	None	$log\ (P_{ij}) = \mu_j + ^{\gamma} Y_{ij}$
M_{th}	time heterogeneity	None	$log\ (P_{ij}) = \alpha_i + \mu_j$
M_{tbh}	time behaviour heterogeneity	None	$log\ (P_{ij}) = \alpha_i + \mu_j + ^{\gamma} Y_{ij}$

Relation of the model
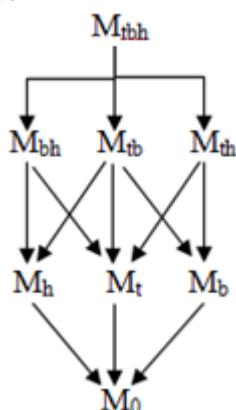
Each arrow points from one model to an immediate special case of the model. Test of assumptions are based partly on these relations among the model
Figure 2: Relation of the Model

There are also capture-recapture models available for the open population like Jolly-Seber model, Modern open population model, Multi State Model, Target Recovery Model and Joint modeling of live and dead recovery, with different assumptions. These models are widely used by researchers to count the animal populations as well as human also.

Log-Linear Models

The log-linear model approach was originally proposed by Fienberg (1972) and has been widely used in many applications (Cormack 1989; Agresti 1994; Evans, Bonett, and McDonald 1994; Coull and Agresti 1999; IWGDMF 1995). In this approach, the data are regarded as a form of an incomplete 2^t contingency table (t is the number of lists) for which the cell corresponding to those individuals uncaptured by any of the samples is missing. Then various log-linear models are fitted to the observed cells and the chosen model is projected onto the unobserved cell by assuming that there is no t-sample interaction. We use the three-sample case as an illustration. There are seven observed cells: z_{111}, z_{110}, z_{101}, z_{100}, z_{011}, z_{010}, and z_{001}, where z_{100} is the number of individuals captured in sample 1 only and z_{011} is the number of individuals caught in samples 2 and 3 but not in sample 1. A similar interpretation pertains to other capture histories. The missing cell z_{000} denotes the number of missed individuals. The log-linear approach models the logarithm of the expected value of each observable category. The most general log-linear model for the three-sample case is

$$log\ E(z_{ijk}) = u + u_1 I(i = 1) + u_2 I(j = 1) + \underline{u}_3 I(k = 1) + u_{12} I(i = j = 1) + u_{13} I(i = k = 1) + u_{23} I(j = k = 1) + u_{123} I(i = j = k = 1);$$
------------------------(1)

where I(A) is an indicator function for event A. The above is a reparameterization of the eight expected values. We have seven observed cells, whereas there are eight parameters in the right-hand side of equation (1). Therefore, a commonly adopted assumption is that there is no three-sample interaction term, i.e., $u_{123} = 0$, which allows an extrapolation formula for the missing cases using the fitted values of the observable cells. The independent model with main effect terms is equivalent to model M_t in Table 1. The log-linear models corresponding to the multiplicative models M_b and M_{tb} are discussed in Cormack (1989), who also showed how to t various models using the GLIM package (GLIM is a statistical package(marketed by NAG) which analyses generalised linear models by iterated reweighted least squares). Many authors stated that random sampling is required in capture studies. This assumption stems from ball and urn experiments, in which marked and unmarked balls are shaken completely, and a random sample is taken at the end of each sampling occasion. Traditional sampling methods include procedures for drawing random samples. Uses of the procedures require knowledge of the sampling probabilities (for finite populations). Deliberate control over the elements to be sampled is required. It is unrealistic to think that if any COVID patient is captured (marked) in one corner of the trapping grid may be captured subsequently in the opposite corner of the boundary. The concept of random sampling does not apply to situations assumed in Models M_b, M_h, M_{bh}, M_{tb}, M_{th}, and M_{tbh}. The goal of these models is to provide an analysis of the sampled data in the face of behavioural response and heterogeneity, both of which are contrary to the traditional role of random sampling.

4. Data Source For Illustration

According to ICMR the data of the following people should be considered for COVID-19 testing.
1) Symptomatic (fever, sore throat, running nose, dyspnea, etc) individuals who have travelled to affected countries

(China, Hong Kong, Japan, South Korea, Singapore, Iran, Italy   etc.)

2) Close contacts of confirmed positive cases of COVID-19 infection.

3) All individuals evacuated and quarantined from Wuhan, China and Diamond Princess ship, Japan.

It includes information on COVID-19 service activities by private and public hospitals and community health services from 25th March 2020 to 31st May 2020 (Lock-down period in India). This study may be use to estimate the number of people who were infected by the corona virus (COVI-19) in an outbreak in whole India and for more accuracy study area may be carried out at state wise. For example if the study area is Delhi selected with approximately 16787941 population in 11 different districts. Our data are restricted to census records of government. In total, 271 cases were reported from the following three sources. (1) A-list: records based on Symptomatic (fever, sore throat, running nose, dyspnea, etc) individuals who have travelled to affected countries (China, Hong Kong, Japan, South Korea, Singapore, Iran, Italy etc.) . There were 135 ascertained cases. (2) B-List: records provided by Close contacts of confirmed positive cases of COVID-19 infection. The records included 122 cases reported by doctors in local hospitals. (3) C-list: records based on All individuals evacuated and quarantined from Wuhan, China and Diamond Princess ship, Japan.. A total of 126 cases were in the C-list.

Table 2: Data on corona virus infection

A	B	C	Data
1	1	1	$Z_{111} = 28$
1	1	0	$Z_{110} = 21$
1	0	1	$Z_{101} = 17$
1	0	0	$Z_{100} = 69$
0	1	1	$Z_{011} = 18$
0	1	0	$Z_{010} = 55$
0	0	1	$Z_{001} = 63$
0	0	0	-----

The data are given in Table 2 Several log-linear models were fitted to these data. Except for the saturated model, the log-linear models, which do not take into account heterogeneity (i.e. models (AC, B), (BC, A), (AB, C), (AB, BC), (AB, AC) and (BC, AC)), do not fit the data well. The quasi-symmetric and three partial quasi-symmetric models (see Darroch *et al*., 1993), which incorporate heterogeneity effects, fit the data well. The four adequate models produce approximately the same estimates, 1300, with an approximate estimated s.e. of 520. This relatively large estimated i.e. shows the data are actually insufficient to model the heterogeneity effect.

Estimation Procedure

$$\hat{C} = 1 - \frac{1}{t} \sum_{k=1}^{t} \frac{s_k}{n_k}$$

Equation for estimator of sample coverage --------------(2)

The singletons cannot contain any overlap information. Therefore, the sample coverage can be Interpreted as a measure of overlap. When neither dependence exists, it follows from the definition of *C* that $C = D/N$, where *D* denotes the average of overlapped cases, i.e.

$$D = \frac{1}{t} \sum_{k=1}^{t} \left\{ \sum_{k=1}^{N} I \left[\sum_{j \neq k} X_{ij} > 0 \right] \right\} = M - \frac{1}{t} \sum_{k=1}^{t} S_k$$

-----------------------(3)

Here, *M* denotes the number of different cases identified in at least one of the lists. Hence, in the independent case, a valid estimator is

$$N_0 = D/\hat{C}$$ --------------------------------(4)

the sample coverage is estimated to be $\hat{C} = 51.27\%$ and the average of the overlapped cases is $D = 208.67$. If we incorrectly assume that the three samples are independent, then from (4) an estimate would give $N_0 = D/\hat{C} = 407$, which is slightly larger than the estimate of 388 based on an independent log-linear model. We feel these data, with estimated sample coverage of 51%, do not contain enough information to correct for undercount. Thus, this example presents a very valuable data set with the advantage of a known true parameter.

5. Discussion

We discuss first the extent to which this study provides evidence for or against the use of mark/recapture techniques in the epidemiological setting. We then consider the relationship of these findings to other similar studies. To what extent then does the present study meet the "requirements" for the mark/ recapture method? As far as possible, the study has attempted to define a "closed" population—Otis condition "a". Note here that "closed" does not infer that the COVID-19 patient population of Delhi is unchanged—rather that the reference population for individual cases is the same for both lists. In this study the reference population is COVID-19 patients from Delhi addresses with following people should be considered for COVID-19 testing.

1) Symptomatic (fever, sore throat, running nose, dyspnea, etc) individuals who have travelled to affected countries (China, Hong Kong, Japan, South Korea, Singapore, Iran, Italy etc.)

2) Close contacts of confirmed positive cases of COVID-19 infection.

3) All individuals evacuated and quarantined from Wuhan, China and Diamond Princess ship, Japan.

Indian Council of Medical Research (ICMR) affirms, while there is no readily available test kit to detect coronavirus infection, doctors and researchers go by a series of tests to confirm the infection i.e. Swab test, Nasal aspirate and tracheal aspirate test. Indian Council of Medical Research (ICMR) suggests that anyone with a recent travel history to COVID-19 affected countries or someone who has been in close contact with infected patients is also advised screening. Despite the extended time over which our lists are compiled, an individual's availability for both lists is almost concurrent. The crucial assumption, that an individual in the population from which one list is a sample, is also in the population from which the other list is a sample, is thus

satisfied. In any area of epidemiology or, in fact, any human monitoring system involving population estimates of the number of people who have a certain characteristic, capture-recapture should be considered a complementary or an alternate approach to traditional counting technology. One application of capture-recapture, in addition to generating an estimate of population size, is to assess the costs of ascertainment relative to the degree of accuracy obtained. A primary goal of monitoring is the ability to obtain accurate and timely incidence rates at a reasonable cost. The more expensive it is to accurately estimate the counts of disease, the less likely it is to assess that disease and the smaller the geographic coverage. When setting out to determine incidence rates, therefore, one should first determine the amount of precision desired, as greater degrees of precision are likely to be more expensive. The appropriate degree of precision required for estimating incidence rates will vary by the use(s) to which the data are put. There are no right or wrong decisions as to how many samples should be examined, or the degree of ascertainment needed, as this is a function of both the acceptable degree of precision and the amount of resources available to carry it out.

We strongly believe that with any disease monitoring system, whether communicable or non-communicable, there should be the means to formally evaluate the degree of undercount. With capture-recapture, we have this technology. All disease monitoring systems should be able to estimate the proportion of the diseases diagnosed in the community, region, or nation that were identified by the monitoring system. Without this information, it is difficult, if not impossible, to determine the degree to which the rates of disease are compromised by missing cases. Capture-recapture is not a panacea. However, it offers public health officials the opportunity to get closer to the truth if the methods are used appropriately. There is much to be learned in the development of capture-recapture models for use in epidemiology. The assumptions need to be assessed and the models tested. Epidemiologists and biostatisticians working with human disease need to continue the dialogue with those who have developed the approaches using capture-recapture. Much remains to be learned, and there is great potential. As indicated by the IWGDMF (1995), the log-linear model approach has many advantages as follows: (a) all models are under a unified framework; (b) model selection can be easily implemented and carried out in a flexible fashion, based on data and guided by prior information; (c) tests are available for comparing models; (d) dependence can be incorporated by adding proper interactions; and (e) all inference is within the mainstream of statistical data analysis. An untestable assumption involved is that the highest-order interaction does not exist. The IWGDMF commented that the existence of heterogeneity in three-list data might result in the lack of a reliable estimate. Another concern is that two equally fitted models might produce quite different estimates. As the number of lists increases, the number of adequate models increases rapidly and thus model selection causes further problems.

References

[1] Anderson, D. R., and Wilson, K. R., 1989: Capture-recapture population estimation using continuous time models. Progress Report.

[2] Becker, N. G., 1984: Estimating population size from capture-recapture experiments in continuous time. Austral. J. Statist. 26, 1-7.

[3] Becker, N. G., and Heyde, C. C., 1990: Estimating population size from multiple recapture experiments. Stochastic Process Appl. 36, 77-83.

[4] Bird SM, King R., "Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy", *Annual review of statistics and its application. 2018 Mar; 5: 95-118,* doi: 10.1146/annurev- statistics-031017-100641

[5] Begon, M. (1983). Abuses of Mathematical Techniques in Ecology: Applications of Jolly's Capture- Recapture Method. Oikos, 40(1), 155. doi:10.2307/3544213

[6] Coeli, C.M., Veras, R.P. and Coutinho, E.S.F., 2000. Metodologia de captura-recaptura: uma opção para a vigilância das doenças não transmissíveis na população idosa. Cadernos de Saúde Pública, vol. 16, no. 4, pp. 1071-1082. PMid:11175530. http://dx.doi.org/10.1590/S0102-311X2000000400025.

[7] Cormack, R.M., 1992. Interval estimation for mark-recapture studies of closed populations. Biometrics, vol. 48, no. 2, pp. 567-576. PMid:1637979. http://dx.doi.org/10.2307/2532310.

[8] Cormack, R.M., 1989. Log-linear models for capture-recapture. Biometrics, vol. 45, no. 2, pp. 395-413. http://dx.doi. org/10.2307/2531485.

[9] Fernandez, F.A.S., 1995. Métodos para estimativas de parâmetros populacionais por captura, marcação e recaptura. In: P.R. PERES-NETO, J.L. Valentin and F.A.S. Fernandez, eds. Programa de pós-graduação em ecologia. Rio de Janeiro: Instituto de Biologia, UFRJ, pp. 1-26. http://dx.doi.org/10.4257/ oeco.1995.0201.01.

[10] Fienberg, S. E. (1972) The multiple recapture census for closed populations and incomplete 2^k contingency tables, Biometrika, 59, pp. 591-603.

[11] Hook, E. B., Albright, S. G., & Cross, P. K. (1980). Use of bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York state. American Journal of Epidemiology, *112(6), 750– 758.* doi:10.1093/oxfordjournals.aje.a113047

[12] H. Brenner (1994). Application of Capture-Recapture Methods for Disease Monitoring: Potential Effects of Imperfect Record Linkage. Methods Inf Med 1994; 33(05): 502-506. DOI: 10.1055/s-0038-1635053.

[13] International Working Group for Disease Monitoring and Forecasting (1995). Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Human Diseases. American Journal of Epidemiology, 142(10), 1059–1068. doi:10.1093/oxfordjournals.aje.a117559

[14] Jones H.E., Welton NJ, Ades AE, Pierce M, Davies W, Coleman B, Millar T, Hickman M. Problem drug use prevalence estimation revisited: heterogeneity in capture–recapture and the role of external evidence.

Addiction (Abingdon, England). 2015 Dec 28; 111(3): 438-447. doi: 10.1111/add.13222

[15] Jarvis SN, Lowe PJ, Avery A, *et al.*Children are not goldfish—mark/recapture techniques and their application to injury data Injury Prevention 2000;**6:**46-50.

[16] Leclerc P, Vandal AC, Fall A, Bruneau J, Roy É, Brissette S, Archibald C, Arruda N, Morissette C.,*"Estimating the Size of the Population of Persons Who Inject Drugs in the Island of Montréal, Canada, Using a Six-Source Capture-Recapture Model"*, Drug Alcohol Depend. 2014 Sep 1; 142:174-80. Epub 2014 Jun 24. doi: 10.1016/j.drugalcdep.2014.06.022.

[17] Nigel Fisher, Stuart W Turner, Robert Pugh, Colin Taylor, Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis , BMJ. 1994 Jan 1; 308(6920): 27–30.doi: 10.1136/bmj.308.6920.27

[18] Robles, S. C., Marrett, L. D., Aileen Clarke, E., & Risch, H. A. (1988). An application of capture-recapture methods to the estimation of completeness of cancer registration. Journal of Clinical Epidemiology, 41(5), 495–501. doi:10.1016/0895-4356(88)90052-2

[19] Sekar, C. & Deming W. E. (1949) On a method of estimating birth and death rates and the extent of registration, Journal of the American Statistical Association, 44, pp. 101-115.

[20] Wittes, J. T. & Sidel, V. W. (1968) A generalization of the simple capture-recapture model with applications to epidemiological research, Journal of Chronic Diseases, 21, pp. 287-301.

[21] Wittes, J. T. (1974) Applications of a multinomial capture-recapture method to epidemiological data, Journal of the American Statistical Association, 69, pp. 93-97.

[22] White, G.C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982) Capture-recapture and removal methods for sampling closed populations. Los Alamos National Laboratory Publication LA-8787-NERP. Los Alamos, NM.

[23] White, G.C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982) Capture-recapture and removal methods for sampling closed populations. Los Alamos National Laboratory Publication LA-8787- NERP. Los Alamos, NM.

[24] Wittes, J. T., Colton, T., & Sidel, V. W. (1974). Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases, 27(1-2), 25–36*. doi:10.1016/0021-9681(74)90005-8

Author Profile

Rakesh Kumar (rakeshtech1@gmail.com) Graduated in Industrial Engineering and Master's in Mechanical Engineering with specialization in Computer Integrated Manufacturing from India. He also obtained Post Graduate Diploma in IPR. He is currently an Assistant Professor and research scholar at Delhi Technological University in Department of Mechanical. His research interest includes Total Quality Management, Reliability Engineering, Operations and Supply chain management, Fuzzy Logic, AI and Decision Science and Intellectual Property Rights. http://orcid.org/0000-0001-5060-1259

R.S. Mishra (Professor) (rsmishra@dce.ac.in) received the Ph.D. degree in Mechanical engineering from IIT Delhi, Indiaand currently at Delhi Technological University serving since 1997. He having 150 research papers in peer reviewed in International journals and 75 research papers in proceedings of International and National conferences. Prior to joining DTU, he served 09 years at different engineering colleges of India, viz, REC Hamirpur, PAU Ludhiana (Punjab)and HAU, Hisar (Haryana.). His research area includes Green Technology, Refrigeration & Air Conditioning, Metal Cutting Technology, Total Quality Management, Solar Energy Technology, Bio Fuels, power Plant engineering.