A REVIEW OF METHODOLOGICAL STRATEGIES FOR ESTIMATING THE TOTAL NUMBER OF EVENTS WITH DATA FROM MULTIPLE-RECORD SYSTEMS

by

M. Nabil El-Khorazaty
Carolina Population Center
University of North Carolina, Chapel Hill

Peter B. Imrey
University of Illinois College of Medicine, Urbana

Gary G. Koch and H. Bradley Wells
Department of Biostatistics
University of North Carolina, Chapel Hill

Institute of Statistics Mimeo Series No. 1095

November 1976
A REVIEW OF METHODOLOGICAL STRATEGIES FOR ESTIMATING THE TOTAL NUMBER OF EVENTS
WITH DATA FROM MULTIPLE-RECORD SYSTEMS

M. Nabil El-Khorazaty
Carolina Population Center, University of North Carolina, Chapel Hill

and

Peter B. Imrey
University of Illinois College of Medicine, Urbana

and

Gary G. Koch and H. Bradley Wells
Department of Biostatistics, University of North Carolina, Chapel Hill

Summary

Two techniques for estimating the total number of events are reviewed in this paper. Through multiple recording of the same event (individual or animal), the Multiple-Record System (MRS) technique (used mainly for human populations) and the Capture-Mark-Recapture (CMR) technique (used mainly for animal populations) attempt to adjust for the incomplete coverage of single systems. The Dual-Record System (DRS) technique, as a special case of the MRS, has been used rather widely to adjust for omissions in the recording of vital events.

The estimation procedure developed for the MRS and CMR have certain limitations because of their inherent assumptions and these may seriously affect the estimates obtained. The use of a log-linear model analysis for incomplete contingency tables, arising from MRS-CMR data, as a methodological strategy for estimating the total number of events, allows choosing an estimation procedure realistically adapted to the properties of actual recording sources of information. Moreover, the incorporation of source correlation and/or event correlation into the estimation procedure can provide insight into the effects of such factors and the strengths and weaknesses of the statistical information systems which have been implemented.

On the other hand, application of the theory already developed for the CMR technique to human populations, in the presence of source correlation, yields more refined estimates of the population size. Comparison of the MRS and CMR in terms of their assumptions and modes of application are given in order to clarify their similarities and differences.
A REVIEW OF METHODOLOGICAL STRATEGIES FOR ESTIMATING
THE TOTAL NUMBER OF EVENTS WITH DATA FROM MULTIPLE-RECORD SYSTEMS

by

M. Nabil El-Khorazaty, Peter B. Imrey, Gary G. Koch, and H. Bradley Wells

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2. THE DUAL-RECORD SYSTEM (DRS) TECHNIQUE</td>
<td>5</td>
</tr>
<tr>
<td>2.1. Survey of Dual-Record Systems</td>
<td>5</td>
</tr>
<tr>
<td>2.2. Effects of Over-Coverage Bias</td>
<td>8</td>
</tr>
<tr>
<td>2.3. Effects of Response Correlation Bias</td>
<td>10</td>
</tr>
<tr>
<td>2.4. Effects of Matching Rule (record linkage) Bias</td>
<td>14</td>
</tr>
<tr>
<td>2.4.1. Structure of the problem</td>
<td>14</td>
</tr>
<tr>
<td>2.4.2. Theoretical studies</td>
<td>16</td>
</tr>
<tr>
<td>2.4.3. Practical studies</td>
<td>17</td>
</tr>
<tr>
<td>3. CONTINGENCY TABLE LOG-LINEAR MODEL ANALYSIS</td>
<td>20</td>
</tr>
<tr>
<td>3.1. Introduction</td>
<td>20</td>
</tr>
<tr>
<td>3.2. Log-Linear Model Formulation</td>
<td>22</td>
</tr>
<tr>
<td>3.2.1. Quasi-independence models</td>
<td>22</td>
</tr>
<tr>
<td>3.2.2. General model designs</td>
<td>25</td>
</tr>
<tr>
<td>3.2.3. Stratification</td>
<td>28</td>
</tr>
<tr>
<td>3.3. Strategies For Fitting Log-Linear Models</td>
<td>30</td>
</tr>
<tr>
<td>4. THE CAPTURE-MARK-RECAPTURE (CMR) TECHNIQUE</td>
<td>35</td>
</tr>
<tr>
<td>4.1. Definition and Assumptions</td>
<td>35</td>
</tr>
<tr>
<td>4.2. Mathematical Modeling</td>
<td>37</td>
</tr>
<tr>
<td>4.3. Epidemiologic and Demographic Applications</td>
<td>42</td>
</tr>
<tr>
<td>5. COMPARING THE MRS (DRS) AND CMR TECHNIQUES</td>
<td>43</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>47</td>
</tr>
</tbody>
</table>
A REVIEW OF METHODOLOGICAL STRATEGIES FOR ESTIMATING
THE TOTAL NUMBER OF EVENTS WITH DATA FROM MULTIPLE-RECORD SYSTEMS

1. Introduction

Throughout the years, researchers in many fields of knowledge have been concerned with the problem of estimating the size of given populations. The elements under investigation might be humans, animals, or events relating to such units. Where counts of organisms are involved, we may be concerned with such quantities as the total number of residents, victims of a congenital anomaly, criminals or victims of crimes, drug abusers, parties involved in automobile accidents, or fish in a lake, deer in a forest, or bacteria on a microscope slide. Populations of events frequently studied are those of births, deaths, migrations, marriages, separations, divorces, and diagnoses of diseases such as cancer within a chosen time period. Many methodological strategies have been developed to deal with this problem of estimating the size and/or growth rate of human and animal populations as well as the incidence rate for various types of vital events.

Censuses (which cover the whole population) and civil registration systems (continuously recording events of interest) supplemented by periodic sample surveys during intercensal periods can supply most of the data required for such estimation purposes for human populations. However, these traditional statistical systems are not adequate in many countries of the world. For obvious reasons, they are non-existent for most animal populations. These limitations of the traditional sources have resulted in the development of statistical procedures using multiple counts of the same event, person or animal, for estimating the number of events or persons and/or adjusting for incomplete coverage of census, survey or vital registration systems. For human populations, these techniques can be generally grouped under the term multiple-record system (MRS)
techniques while for animal populations the methods generally are called capture-mark-recapture (CMR) techniques.

The multiple-record system (MRS) technique involves data collection from two or more sources of information (often called recording systems) which cover the same sample or sub-sample of areas and the same time period. The special case of two sources (known as the dual-record system (DRS)) has been used rather widely in the last thirty years to adjust for omissions in the recording of vital events and to estimate population growth rates. In this regard, Chandrasekar and Deming (1949) present a theoretical framework for estimating the total number of events under the following inherent assumptions:

1.1 There are no coverage errors with respect to the scope of area and/or time period in which events are recorded (i.e., each information source only records events that pertain to the target area and/or time period under study).

1.2 The information sources are independent (i.e., the probability for an event being recorded by one source does not depend on whether it is recorded by the other source).

1.3 There are no misclassification errors with respect to determining whether a particular event has been recorded by both information sources or only one of them (i.e., a perfect matching rule exists for linking the two information sources together in terms of the number of events which are recorded by both).

1.4 There are no recall lapse errors with respect to the recording of events (Som, 1973).

Chakraborty (1963) and Das Gupta (1964) extend this approach to situations involving three or more sources of information.

The Capture-Mark-Recapture (CMR) technique has been used for more than three centuries for estimating the total size of animal populations (Marks, Seltzer and Krotki (1974)). In the simplest (and classical) formulation, the investigators initially capture and "mark" as many animals as feasible, subsequently returning them to the wild. The "mark" may be a tag on the foot, a colored band, a brand or any other unequivocally artificial and identifiable
attachment or alteration. Sometime later, as many animals as possible are observed, with a ratio of marked among these to total marks used to obtain a population size estimate. Peterson (1896) and Lincoln (1930) obtained estimates in this manner; their "Peterson-Lincoln index" (P-L index) precedes and is equivalent to the Chandrasekar-Deming (C-D) formula for the dual-record system technique. More recently, other deterministic as well as stochastic models have been developed for dealing with this estimation problem. These more general models are described by Seber (1965), Jolly (1965), Robson (1969), Pollock (1975), and El-Khorazaty and Sen (1976).

A fruitful approach to the study of MRS and CMR data is to view the involved set of detection protocols (recording systems for MRS, capture or observation methods or times for CMR) as a probabilistic process, or channel, with an input and resulting output. Input to the processor is a single element of the population to be studied, while the resulting output is a response pattern which delineates exactly which of the various detection protocols (if any) have recorded, or captured, that element. The aggregate result of passing every population element through the processor may be arranged, for d detection protocols, in a $2^d$ contingency table with each dimension, or marginal, describing the success or failure of a single detection protocol in recording the elements of the target population. In a typical MRS or CMR application, we observe this contingency table except for the single cell containing those elements missed by all detection protocols. It is this "incomplete contingency table" which must be used to generate estimates of population size. More generally, the population under study sometimes may be partitioned into a set of subpopulations according to such demographic variables as geographic location, urbanization, sex, or age, so that we observe such an incomplete table with its missing cell for each subpopulation.

Fienberg (1972), Bishop, Fienberg and Holland (1975), El-Khorazaty (1975), Koch, El-Khorazaty and Lewis (1977), and El-Khorazaty, Imrey, Koch, and Lewis
(1976) advocate fitting log-linear models to the observed cells of the previously described contingency tables, and using these models to obtain refined estimates of the missing cell(s) and, hence, of the population total. Such refined estimates may be obtained by

i) controlling for statistical dependence of specific types among the actions of the various detection protocols;

ii) accounting for or modeling the effects of subpopulations or their defining factors on the probabilistic properties of the detection protocols.

This strategy, which generalizes the Chandrasekar-Deming approach for the DRS and the Peterson-Lincoln approach for the CMR, allows great latitude in choosing an estimation procedure, realistically adapted to the properties of actual recording systems.

This paper reviews the MRS-CMR problems, with particular reference to the effects of violations of the Chandrasekar-Deming Assumptions (1.1)-(1.4) and to recently developed methods which avoid or ameliorate such effects. These methods are, perhaps, most relevant to developing countries, as they can provide insight into the strengths and weaknesses of statistical information systems which are implemented.
2 The Dual-Record System (DRS) Technique

2.1 Survey of Dual-Record Systems

The idea underlying the DRS estimation procedure is simple and was well-known in the early days of civilization as the "rule of three." If

\[(a/b) = (c/d)\]  \hspace{1cm} (2.1)

and any three of \(a, b, c, d\) are known, the fourth can be determined. Assume that two different sources of information (systems) are operating to detect each of the \(N\) elements of a target population. Let \(n_g\) \((g=1,2)\) represent the total numbers of elements detected by source \(g\), and \(n_{11}\) be the number detected by both. Table I shows these quantities in the contingency table format described in Section 1. If detection by Source II is unrelated to detection by Source I, as in (1.2), then the proportion of the population detected by Source II should equal the proportion that Source II detects in the subpopulation of those elements detected by Source I. Hence, ideally, \((n_{11}/n_1) = (n_2/N)\). If we "make this true" with the rule of three, we have

\[\hat{N} = n_1 n_2 / n_{11}\]  \hspace{1cm} (2.2)

the estimate proposed by Chandrasekhar-Deming (C-D) and Peterson-Lincoln (P-L).

The corresponding estimate for the lower-right cell, the number of elements missed by both sources, is \(n_{12} n_{21} / n_{11}\). Chandrasekar and Deming (1949) give a
mathematical formulation which justifies $\hat{N}$ under assumptions (1.1)-(1.4); we discuss derivation further in Section 3.

Although the concept of a dual-record system technique had been known for many years, Coale (1961) stimulated demographic application by suggesting that civil or special recording sources be supplemented by periodic sample surveys to form dual-record systems for the assessment of vital rates in human populations. Since then, such systems have been formed in many countries and used to study population growth rates and/or to evaluate the coverage of civil systems.

Linder (1970) gives an extensive description of the DRS technique and its limitations, while Abernathy and Lunde (1970) discuss early history and give detailed presentations of applications in different countries. The primary use of the DRS has been to estimate rates of birth and death, which are "rare" events. According to Wells (1971), in this context DRS "can provide better estimates and also provide a framework for evaluating different single systems." Lauriat (1967) gives a comparative study of DRS applications in Pakistan, Thailand and Turkey, while Seltzer (1969) discusses applications in India as well as additional topics. The administrative setting for these countries and Liberia is described in Wells (1971). The Agency for International Development stimulated the establishment of an International Program of Laboratories for Population Statistics (POPLAB) to "establish long-range cooperative programs of work with institutions in various countries," and to "improve the science of collecting and analyzing demographic statistics" (Linder (1971)). Through this program, dual-record systems in Colombia, Philippines, Liberia, Morocco, Turkey, Kenya, and Ecuador relate to a central POPLAB research component, and to one another. On the other hand, questions about the usefulness of the DRS technique in estimating events missed by both sources are raised by Brass (1971, 1975), who argues that because of the independence assumption, the
DRS detects only a small part of the omissions. He also states that the technique is very costly because of the need for preservation of independence, matching, and efficient administrative organization. Thus, he suggests using the easier and more economical single system designs that rely on repeated surveys. Finally, the most comprehensive work on the DRS, its background, assumptions, design, and problems is the handbook by Marks, Seltzer and Krotki (1974), while a recent bibliography on the subject is published by Carver (1976).

The DRS estimation procedures depend on data obtained through two different information sources - usually a continuous recording system and a periodic survey. There are four different sample design situations:

1. both the first and the second sources cover the whole population;

2. the first and the second sources cover the same sample of the whole population, as Coale (1961) proposed;

3. the first source covers the whole population while the second covers a sample of it;

4. the first source covers a sample of the whole population while the second one covers another sample of it, and the two samples have some areas in common.

While the first two situations pertain to the same area, the third considers one of the sources as a sub-sample of the area covered by the other (census or registration system), and the fourth is of interest for overlapping areas. Data analysis in situations 2-4 may vary according to whether probabilistic properties of the sampling scheme are considered as internal or external to the recording system mechanism.

When a sample survey has been used as Source II, the selection procedure has usually been area sampling of clusters of households, with a single household as the unit of analysis. As with any population survey (including census), all sources of error should be thoroughly examined. With respect to dual-record systems, this means that each component source must be singly studied
to assess both sampling and response errors which may bias the component; additionally, sources of error for each component should be examined jointly in terms of their combined effect on the population size estimator \( \hat{N} \).

Marks (1969, 1971) discusses three components of the variance of \( \hat{N} \):

1. "sampling" variance, due to the probabilistic nature of the sampling and detection process;
2. "non-sampling" variance, due to response errors in registering population elements;
3. "matching" variance, due to errors in the process by which we identify those elements captured by both systems.

In any DRS technique, the three sources of variance are interrelated, particularly non-sampling and matching errors. Sampling variance may be reduced by improved stratification, reduced clustering and increased sample size in the survey component of the DRS. The last two modifications may also decrease the other variance components, but at increased cost.

With respect to bias, Seltzer and Adlakha (1969) show that biases introduced by over-coverage (spurious events), "response correlation," or matching are of particular importance with respect to the statistical properties of the estimate \( \hat{N} \). In the rest of this section, we discuss the nature of bias from these sources as based on their study and those of Jabine and Bershad (1968), Marks (1969, 1971), Seltzer (1969, 1971, 1972), Marks, Seltzer and Krotki (1974), Greenfield (1975), and Greenfield and Tam (1976).

2.2 Effects of Over-Coverage Bias

The estimator \( \hat{N} \) is subject to systematic errors induced by deviations of the populations accessed by each single record system from the target population. The coverage bias, in general, is the excess of the over-coverage over non-coverage bias, or vice versa. By over-coverage, we mean the inclusion of events that do not belong either to the area or to the time period specified. By non-coverage,
we mean the omission of events that belong both to the area and to the time period specified. Basically, the estimation procedure of the DRS is "a method of correcting one form of coverage error - erroneous omissions" (Marks, Seltzer and Krotki (1974)). Thus, the DRS technique attempts to minimize the extent of non-coverage errors. However, the problem of over-coverage errors must still be controlled. Over-coverage bias can arise because of infringement into the wrong areas (e.g., by going beyond boundaries) and/or the telescoping of events forward or backward in time. Boundary errors will be related to the size of the areas used for matching, where usually the smaller the area the greater the chance of boundary errors.

Marks, Seltzer and Krotki (1974) states that over-coverage bias "operates much as it does for estimates based on a single source. That is, the inclusion of events which occurred outside the specified time or geographic limits will bias the estimate upward." However, they add that over-coverage makes a greater contribution to net error in the DRS studies. This over-estimation arises because of the inclusion of too many events in either \( n_1 \) or \( n_2 \) or both. More specifically, if the over-covered areas or time periods for the two sources are assumed to be non-overlapping and if perfect matching is assumed, then

\[
\hat{N} = \left\{ \sum_{g=1}^{2} \frac{(n_g^* + c_g)}{n_{11}} \right\}
\]

(2.3)

where

\[ n_g = n_g^* + c_g \]
\[ c_g = (r_{g1} - r_{g2}) + (t_{g1} - t_{g2}) > 0 \]

\( n_g^* \) is the true number of events that should be recorded in the \( g \)-th source, if there were no area or time error; \( r_{g1}, t_{g1} \) are the over area and time errors for the \( g \)-th source, and \( r_{g2}, t_{g2} \) are the under area and time error for the \( g \)-th source (\( g=1,2 \)). In case of overlapping in coverage errors (either area or time period), the effects on the estimation procedure are complicated by the sensitivity of \( \hat{N} \) to the extent of overlapping of the individual constituent sources.
Finally, within the framework of (2.3), under-estimation of the total number of events may occur with the C-D formula when \( n_1 \) or \( n_2 \) or both include too few events. This type of problem may arise for the third and fourth sample design situations discussed in Section 2.1. Though the combined effects of coverage errors can thus bias \( \hat{N} \) in either direction, in practice over-coverage and net over-estimation are considered dominant.

To minimize such bias Chandrasekar and Deming (1949), Coale (1961), and Seltzer and Adlakha (1969) suggest a field re-investigation for non-matched events to detect and eliminate events wrongly recorded. Seltzer and Adlakha (1969) recommend the use of concepts "that are unambiguous and meaningful." Marks (1971) suggests, if both sources of information are samples, "larger sampling units (with well-defined boundaries) should be selected for one of the data collection sources and one-way matching [which determines the matching status of all the events of only one source] applied over the entire area." Over-coverage bias may also arise from double-counting of an event by a single information source, but in the DRS this bias can be eliminated by special attention to accurate record-management.

### 2.3 Effects of Response Correlation Bias

Response correlation bias refers to a limitation of the C-D (P-L) estimator due to the failure of assumption (1.2). Table II displays conceptual probabilities corresponding to the four possible results of joint operation of the dual-record system technique on a single element selected randomly from the population. Then (1.2) is equivalent to

\[
\pi_{j_1j_2} = \pi_{j_1} \pi_{j_2} \tag{2.4}
\]

or alternatively,

\[
\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1. \tag{2.5}
\]

The cross-product (or odds) ratio can, of course, be equal to one only in
theory; in application numerous specific characteristics of the recording systems and the population produce departures from this model, and if the net departure is large, the C-D (P-L) estimator is severely compromised.

Table II. Detection Probabilities of a Dual-Record System Technique

<table>
<thead>
<tr>
<th>SOURCE II</th>
<th>DETECTED</th>
<th>UNDETECTED</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETECTED</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$\pi_1$</td>
</tr>
<tr>
<td>SOURCE I</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$\pi_2$</td>
</tr>
<tr>
<td>UNDETECTED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>$\pi_{1}'$</td>
<td>$\pi_{2}'$</td>
<td>1</td>
</tr>
</tbody>
</table>

Underlying causes of response correlation are described by El-Khorazaty (1975) in terms of the concepts of "source" and "event" correlations. Source correlation is present whenever the detection probability of a particular population element with respect to Source II changes if we condition on the result of the Source I screen. This consideration pertains to the operating characteristics of the sources of information with respect to single elements, rather than the average operating characteristics referred to in (1.2) and Table II. Source correlation exists, then, whenever the odds ratio

$$\frac{\pi_{11}(k)\pi_{22}(k)}{\pi_{12}(k)\pi_{21}(k)} \neq 1$$  \hspace{1cm} (2.6)

for any $k$, where $\pi_{j_1j_2}(k)$ is the probability that population element $k$ arrives in the $(j_1, j_2)$-cell of Table I after processing by the DRS. If the population is homogeneous, $\pi_{j_1j_2}(k) = \pi_{j_1j_2}$ for all $j_1, j_2$, and $k$. On the other hand, event correlation arises when the population is heterogeneous with respect to the operating characteristics of the DRS, in which case the overall odds ratio, $(\pi_{11}π_{22}/π_{12}π_{21})$, will ordinarily not equal one even if all the element-specific
ratios \( \pi_{11}(k)\pi_{22}(k)/\pi_{12}(k)\pi_{21}(k) = 1 \). This is easily seen using

\[
\pi_{j_1 j_2} = \frac{1}{N} \sum_{k=1}^{N} \pi_{j_1 j_2}(k) \tag{2.7}
\]

and virtually any set of \( \pi_{j_1 j_2}(k) \)'s. Population heterogeneity thus produces violations of (1.2) even when the two components of the DRS are stochastically independent in their processes with respect to each element singly. Apparent correlation between components of the DRS, produced indirectly through differential operation of the DRS on the elements of the population of individuals or events, is termed event correlation.

To clarify the distinction, note that source correlation typically arises through (Seltzer and Adlakha (1969)):

i) planned communication between Sources I and II, such as the use of the same or similar maps, instructions, or numbering systems; or

ii) unplanned communication such as informal collaboration between personnel of Sources I and II.

Event correlation, on the other hand, occurs through failure of each source to have equal access to all subgroups of the population. Jabine and Bershad (1968) indicate that "it is reasonable to believe that for certain classes [of the population], the events would be more difficult to observe or to record than for others, regardless of the method used." Hence, as mentioned by El-Khorazaty and Sen (1976), "dependence is inherent and it affects the estimation process."

Response errors typical of individual surveys undoubtedly exert primary effects in terms of event correlations, but may also produce source correlation. Hansen, Hurwitz, Marks and Mauldin (1951) provide a general discussion of response errors, in the context of models for total survey error developed by the U.S. Bureau of the Census. Among contributors to response errors are:

A. Errors Attributable to Respondent

1. Responses supplied by different members of household;
2. Variation in response of some individuals over time, due to
   a) failure of memory (Som, 1973), inadvertent reporting mistakes, misunderstanding of questions, confusion due to
take change in survey procedures (Deming (1966));
   b) protection of privacy or prestige, hostility to govern-
   ment, surveys, or interviewer.

B. Errors Attributable to Interviewer

1. Variation in interpretation of questions, definitions, and
   instructions between and within interviewers;

2. Interviewer suggested responses based on relationship with
   respondent.

The response correlation coefficient,

\[ \rho = \frac{\pi_{11} - \pi_{1.} \cdot \pi_{.1}}{\sqrt{\pi_{1.} \cdot \pi_{.1}(1 - \pi_{1.})(1 - \pi_{.1})}}. \]  

(2.8)

is used to measure the extent of dependence between the two components of a

DRS. While negative correlation may arise when the components each tend to
capture or record most effectively different types of events, such correla-
tions in recording vital events have usually been positive in practical situa-
tions. Jabine and Bershad (1968) showed that such positive correlation produces
a downward bias in \( \hat{N} \). Heuristically, the DRS technique systematically misses
more elements than would be expected if the components were independent. Thus,
conditional on the observed data the C-D (P-L) estimate is always too low in such
situations and the unconditional estimator is downward biased. As a result,
\( \hat{N} \) may be regarded as a lower bound for a reasonable estimator. Adding an
additional assumption as to the maximum possible response correlation \( \rho \),

Greenfield (1975) and Greenfield and Tam (1976) derive a similar upper limit
for the missing cell as
\[ \hat{n}_{22}(\text{max}) = \frac{1}{2} \left( B + \frac{2}{\sqrt{B^2 + 4A^2}} \right), \quad (2.9) \]

where
\[ A = \frac{n_{12}n_{21} \left( 3 \sqrt{n_{12}n_{21}} - n_{11} \right)}{\sqrt{n_{12}n_{21}} - 3n_{11}}, \]

and
\[ B = \frac{(n_{12} + n_{21}) \left( n_{11} - \sqrt{n_{12}n_{21}} \right)^2 + 8n_{11} n_{12} n_{21}}{(n_{11} + \sqrt{n_{12}n_{21}}) \left( 3n_{11} - \sqrt{n_{12}n_{21}} \right)}. \]

Approximations for special cases and a table of correction factors for these approximations were also provided. Chandrasekar and Deming (1949) suggest that the effects of dependence could be minimized by partitioning the subject population into essentially homogeneous subpopulations, each with low internal source correlation. Wittes (1970) treats this problem in the MRS context. Even when source correlations preclude such a fortuitous partitioning, the separate analysis of internally homogeneous classes can be used to estimate and adjust for source dependence of some types. Methodology for efficiently using such a subpopulation structure will be discussed in Section 3.

2.4 Effects of Matching Rule (Record Linkage) Bias

2.4.1 Structure of the Problem

Since the components of a dual-record system operate essentially separately and in parallel, a record linkage procedure must be utilized to identify those population elements captured by both components. Choice and implementation of such a procedure is susceptible to many difficulties "because the notion of a 'good match' is undefined" (Okner (1974)). However, the search for the "matching rule" which minimizes the matching error bias is a principal concern of many authors because of its implications to the properties of
the estimator \( \hat{N} \). This is easily seen by writing

\[
\hat{N} = \frac{n_1 n_2}{n_{11}} = \frac{n_1 n_2}{n_{11}^* + e_m - e_{nm}} \tag{2.10}
\]

where \( n_{11}^* \) is the true number of population elements identified by both sources, \( e_{nm} \) is the number of such elements inadvertently not identified by the matching rule (i.e., erroneous non-matches), and \( e_m \) is the number of pairs of records of different elements which the matching rule mistakenly attributes to the same element (i.e., erroneous matches). For a good matching rule, both \( e_m \) and \( e_{nm} \) should be stochastically small. In practice, however, a decrease in one type of error is frequently paid for by an increase in the other. A rule characterized by stochastically large \( e_{nm} \) and small \( e_m \) is said to be "tight," and results in an estimator \( \hat{N} \) which is systematically too large. A rule with small \( e_{nm} \) and large \( e_m \) is said to be "loose," and \( \hat{N} \) then systematically underestimates \( N \). This disposition of matching rules is referred to by some authors (El-Khorazaty, 1975) as the "matching dilemma." We review the literature of record linkage in this section, keeping effects on the C-D (P-L) estimator in mind.

Record linkage was defined by Dunn (1946) as a system with which "to link the various important records of a person's life," such as birth and death records ("death clearance"). The last two decades, encompassing the wide distribution of digital computers, have seen a rapid growth of interest in and literature pertaining to this subject. Wagner and Newcombe (1970) give a bibliography on the methodology of linkage and its applications in medical data processing.

Acheson (1967) describes a two-stage process of record linkage. Initially a chosen record from the first file is compared to all records from
the second file with respect to a limited number of specific identifying characteristics. All records in the second file similar to that from the first file are identified. These records are then each compared in detail to the chosen record and a decision is made as to whether each matches. The process is repeated for each element of the first file. The initial stage, involving many individuals but only a few comparison variables, is called the "search." The second stage, involving the comparison of many variables among only a few individuals, is called the "match" stage. Nitzberg and Sardy (1965) and Acheson (1967) have discussed desirable characteristics of variables to be utilized as criteria in such a procedure. Such identifying characteristics should be permanent and unique for each element, measurable on every element and recorded in both files with high accuracy, and highly discriminating in the sense that single values of the characteristic are shared by only a small number of elements.

Sunter and Fellegi (1967), Tepping (1968) and DuBois (1969) have considered the choice of a matching rule which is in some sense optimum based on the information available. Since appropriate criteria of optimality generally vary with needs and objectives of the linkage, as well as with design of the initial data collection mechanisms, no widely applicable objective theory is available. Tepping (1968) presents a model which allows the minimization of expected total costs associated with implementation of a decision rule; however, the use of this model in practical applications may be difficult.

2.4.2 Theoretical Studies

Several workers have considered the problem of formulating a matching rule based on one identifying variable, when recording of that variable is always accurate and complete. Goodman (1952) and Deming and Glasser (1959)
deal with the two-file situation encountered in operating a DRS. Nathan 
(1967) studies the same situation, but for the purpose of combining infor-
mation from the two files rather than estimating the number of common ele-
ments. He minimizes a cost function defined in terms of appropriate proba-
bilities. Raj (1961) considers sampling from a single file with possibly 
multiple folders on the same individual.

Sunter and Fellegi (1967) use a hypothesis testing approach to matching 
with more than one criterion variable, in the possible presence of recording 
errors and missing observations. Tepping (1968, 1971) assumes no missing 
observations, but otherwise extends their approach by developing a conditional 
utility function of the three matching actions i) definite match, ii) possible 
match, and iii) no match. To apply Tepping's model in practice requires estima-
tion of a loss function and some conditional probabilities. DuBois (1969) uses 
binomially distributed indicator variables to describe missing information and 
the presence of a match. Maximum likelihood is used to estimate the model 
parameters.

DeGroot, Feder and Goel (1971) use an entirely different approach. They 
assume that the joint distribution of the two identifier vectors is multi-
variate normal and develop a method of pairing observed vectors to maximize 
the probability of reproducing an initial set of matches. Yesilcay (1975) uses 
the mean square error (MSE) of the estimate \( \hat{\theta} \) as a criterion for determining 
the "optimum" matching rule (which results in minimum MSE for fixed cost, or 
minimum cost for fixed MSE).

2.4.3 Practical Studies

Applications of record linkage procedures have been widespread and are 
becoming more so. Shapiro (1950, 1954) and Binder (1962) studied under-
registration of U.S. births by matching birth records against census re-
cording of infants and against death records. Birth, death, hospital in-
patient and domiciliary obstetric records have been linked together in medical
research (Acheson (1967)). Bellocco and Arellano (1973) describe the linkage of
records from a 1965 health survey of Alameda County, California to 1965-70
the linkage of child health records in England, Czechoslovakia, and Finland.
Dunn and Gilbert (1956) suggest linking individual records to family records,
and Newcombe, James and Axford (1957) develop such a method. Newcombe,
Kennedy, Axford and James (1959), and Newcombe and Rhynes (1962) link birth
records with marriage records. From their point of view this would be
a powerful tool in the development of population genetics. Recent develop-
ments in genetics to construct a human chromosome map using autosomal linkage
are discussed in Falk, Walker, Martin and Allen (1975). These developments,
as well as the feasibility of computerizing the linkage process, motivated
recent trends to establish extensive national statistical systems (Lunde, 1975a)
in developed and developing countries by bringing together data available from
different sources. For detailed country experiences, see Nordbotten (1967) and
Aurbakken and Bjerve (1973) for Norway, Ohlsson (1967) for Sweden, Steinberg and
Pritzker (1967) and Lunde (1975b) for the U.S.A., Bartels and Fürst (1973) for
West Germany, Donda (1973) for East Germany, Huszár (1973) for Hungary, and
Redfern (1973) for the U.K.

Regarding other applications, Wittes (1970), Wittes, Colton and Sidel
(1974) and El-Khorazaty and Sen (1976) discuss estimating the number of children
in Massachusetts afflicted with a congenital anomaly. In addition, matching of
medical records has been used to estimate the number of infections among hospi-
talized patients (Lewis and Hassanein (1969)) and the number of patients re-
ceiving Methicillin (Wittes (1974) and El-Khorazaty (1975)). Other uses have been for the study of psychiatric admission rates (Pollack (1965)), employee wage records (Binder (1962)), and subscription fulfillment by publishing companies. Applications in sociology, travel studies, education, and economics may be found in Christensen (1958), Davidson (1962), Croxford (1968) and Okner (1974), respectively.

Practical discussions bearing on the choice of matching rule deal either with coverage checks of a census (Shapiro (1950, 1954), Binder (1962), Perkins and Jones (1965) and Hambright (1969)), or with experiences in implementation of a DRS in developing countries (Murty and Jain (1967), Srinivasan and Muthiah (1968) and Mehta (1970) in India as well as Yusuf (1968) in Pakistan). Together, these studies indicate

1. No single matching procedure is optimal over all local variations in designs, objectives, resources, and costs.

2. Most systems have attempted to utilize a rule which
   a. incorporates as many variables with high discriminating power as possible; and
   b. minimizes gross matching error (sum of erroneously matched and erroneously non-matched records).

3. The most satisfactory results in executing matching procedures are obtained through use of computers in the early screening comparisons, and through comparison by humans in the more detailed stages. For such computer-oriented applications, see Newcombe et al (1959), Kennedy (1962), Phillips and Bahn (1963), Nitzberg and Sardy (1965), Newcombe and Kennedy (1965), Nitzberg et al (1967), Belloc and Arellano (1973), Marks et al (1974), Yesilcay (1975), and Falk et al (1975).
3 Contingency Table Log-Linear Model Analysis

3.1 Introduction

In a MRS with $d$ component recording systems, let $n_{j_1 j_2 \ldots j_d}$, $j_g = 1,2$ and $j_g = 1$ for some $g$ represent the number of population elements detected by exactly those recording systems whose corresponding subscripts equal one. The $n_{j_1 j_2 \ldots j_d}$ may be tabularly arranged as a $2^d$ cross-classification with one missing cell; this array contains all the data available for estimating the population size $N$ and constitutes an "incomplete" contingency table, in the sense that the count for one cell of the natural cross-classification is unavailable to us. Such tables have been studied extensively by Goodman (1968), Bishop and Fienberg (1969), Mantel (1970), Fienberg (1972), Haberman (1974) and Bishop, Fienberg and Holland (1975).

Estimation of population size $N$ from MRS (DRS) data has generally proceeded along one of two lines:

i) estimation based on a multinomial model for the distribution of the incomplete table, conditional on the total $n$ of elements detected by any of the record systems;

ii) estimation based on a multinomial model for the distribution of the complete cross-classification, including the unobserved $n_{22\ldots2}$.

Analytic advantages accrue from the first approach, as it allows interpretation and extension of the C-D (P-L) formula in the context of the theory of log-linear models for observed contingency tables. A log-linear model formulation provides a convenient framework to make some adjustments for event correlation by subdividing the study population (and sample) into more homogeneous subgroups, estimating separately for each subgroup such operating characteristics of the recording systems as may differ across them, and reassembling the results. In addition, the approach allows incorporation of such source correla-
tion as may be observable from the data into the estimation procedure, in a manner strictly under control of the analyst. Finally, the results of an analysis based on multiple demographically delineated subgroups may be further examined to study the manner in which operating characteristics of the MRS relate to demographic variables. Given this perspective, further discussion of the structural aspects of log-linear models for DRS and MRS data and statistical methods for fitting such models is given in Sections 3.2 and 3.3.

As stated in the previous paragraph, computational convenience is one of the principal advantages of the conditional approach, particularly for complex (sampling and/or registration system) designs involving several sources. However, the applicability of this approach is also subject to the limitations implied by its inherent assumptions (e.g., multinomial distribution, no coverage errors, no matching errors, no source correlation, no event correlation). Alternatively, the unconditional approach provides a framework for dealing with situations where such assumptions are not realistic, provided that they can be characterized by an appropriate stochastic model. Such models can be formulated for certain types of CMR applications (animal or event populations) as reviewed in Cormack (1968). In this regard, estimation is usually based on maximizing the unconditional likelihood function which may require very extensive computations. Thus, this approach may not necessarily be feasible for the complex (sampling and/or registration) systems associated with MRS applications (human populations). Finally, Sanathanan (1972a, 1972b, 1973) considers conditional estimation of the population size N for the CMR situation and indicates that the conditional and unconditional approaches provide consistent estimators that have the same asymptotic normal distribution. Further discussion of these issues is given in Section 4.
3.2 Log-Linear Model Formulation

3.2.1 Quasi-Independence Models

We now suppose that each of a total of N elements from a homogeneous population presents itself to a dual-record system, which classifies each element independently into one of the cells of Table I according to whether or not the element is detected by either, none or both sources of the system. The state of the system is assumed constant, so that each population element is classified into the four cells of Table I with probabilities \( \pi_{11}', \pi_{12}', \pi_{21}', \pi_{22}' \), respectively, where \( \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \pi_{j_1 j_2} = 1 \), and the \( \pi_{j_1 j_2} \)'s reflect the properties of the dual recording mechanisms (Table II). If the two mechanisms are source-independent, then

\[
\pi_{11}' = \pi_{1}. \pi_{1}. \tag{3.1}
\]

where

\[
\pi_{j_1} = \frac{2}{\sum_{j_2=1}^{2} \pi_{j_1 j_2}} \quad \text{and} \quad \pi_{j_2} = \frac{2}{\sum_{j_1=1}^{2} \pi_{j_1 j_2}}. \tag{3.2}
\]

If N were known, we would estimate the components of equation (3.1) by \( n_{11}/N, (n_{11} + n_{12})/N, \) and \( (n_{11} + n_{21})/N, \) respectively. Replacing \( \pi_{11}', \pi_{1}. \) and \( \pi_{.1} \) by their estimates yields the C-D (P-L) formula for \( \hat{N}. \)

More formally, under the conditions (1.1) and (1.3) the distribution of the observed vector \( \tilde{n}' = (n_{11}', n_{12}', n_{21}') \) is multinomial

\[
P(\tilde{n}) = \frac{n_{j_1 j_2}^{n_{j_1 j_2}}}{\sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \pi_{j_1 j_2}^{n_{j_1 j_2}}}. \tag{3.3}
\]

If the DRS is source-independent, we have
\[
\frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{21}}{\pi_{22}} \quad \text{or equivalently,} \quad \pi_{j_1j_2} = \pi_{j_1} \cdot \pi_{j_2}.
\] (3.4)

By choice of an appropriate constant \( \mu \), the last expression may be written as

\[
\log \pi_{j_1j_2} = \mu + \beta_{j_1} + \beta_{j_2}
\] (3.5)

where

\[
\sum_{j_1=1}^{2} \beta_{j_1} = \sum_{j_2=1}^{2} \beta_{j_2} = 0 \quad \text{and} \quad \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \pi_{j_1j_2} = 1.
\] (3.6)

This is the simplest log-linear model of interest for contingency table analysis. By changing the upper limits of summation on the two restrictions, we obtain models equivalent to the hypothesis of independence of rows and columns in any two-way table. However, the model is not useful in conjunction with the unconditional multinomial distribution given because no estimates of the \( \pi_{j_1j_2} \) are available for use in model fitting. Suppose, instead, that we consider the conditional distribution of \( n \) based on fixed \( n = n_{11} + n_{12} + n_{21} \), the total number of elements detected by either record system. We obtain

\[
P\{n|n\} = \frac{n! \prod_{(j_1,j_2) \neq (2,2)} \pi_{j_1j_2}}{\pi_{j_1j_2}}
\] (3.7)

where

\[
\frac{\pi_{j_1j_2}}{\pi_{j_1} \cdot \pi_{j_2}} = \frac{\pi_{j_1j_2}}{1 - \pi_{22}},
\] (3.8)

the conditional a posteriori probability that a detected element was captured by exactly those recording systems whose corresponding subscripts are one.

The log-linear model then gives
\[
\log \pi_{j_1j_2} = \log \pi_{j_1j_2}^{(1)} - \log (1 - \pi_{22}^{(2)})
= \mu + \log (1 - \pi_{22}^{(1)}) + \beta_{j_1}^{(1)} + \beta_{j_2}^{(2)} = \mu + \beta_{j_1}^{*} + \beta_{j_2}^{*2}
\]

for \((j_1, j_2) \neq (2,2)\). This model structure is called "quasi-independence," since it is the independence structure applied to the distribution of a selected set of cells conditioned on their total. The model is useful because natural estimates are available for the \(\pi_{j_1j_2}\), viz \(n_{11}/n, n_{12}/n, n_{21}/n\), which maximize the conditional likelihood when substituted for \(\pi_{11}, \pi_{12}, \pi_{21}\), respectively. Using these estimates, we may solve by simple algebra for the estimates of \(\beta_{j_1}^{*}\) and \(\beta_{j_2}^{*}\), which maximize the conditional likelihood under the quasi-independence model. Since these parameters are identical to those in the model for independence in the full four-cell table, it is reasonable to use our estimates of them and the model

\[
\log \pi_{j_1j_2} = \mu + \beta_{j_1}^{*} + \beta_{j_2}^{*},
\]

\[
\sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \pi_{j_1j_2} = 1
\]

(3.10)

(3.11)

to estimate all the \(\pi_{j_1j_2}\), \(j_1, j_2 = 1,2\), and then use \(\hat{N}_{QI} = \frac{n}{1 - \pi_{22}}\) to estimate \(N\). The outcome of this procedure gives \(\hat{N}_{QI} = \hat{N}\), the C-D (P-L) estimator.

The previous discussion extends immediately to the general MRS, of order \(d\), for which data are expressible in an incomplete \(2^d\) contingency table with one missing cell. Define \(\pi_{j_1j_2\ldots j_d}\) analogously to \(\pi_{j_1j_2}\). Then the distribution of the \(n_{j_1j_2\ldots j_d}\) is
\[ p \{ n_{j_1j_2\ldots j_d} \} = N! \prod_{j_1=1}^{2} \prod_{j_2=1}^{2} \ldots \prod_{j_d=1}^{2} \frac{n_{j_1j_2\ldots j_d}}{n_{j_1j_2\ldots j_d}} \] (3.12)

and the quasi-independence model

\[ \log \pi_{j_1j_2\ldots j_d} = \mu + \sum_{g=1}^{d} (g) \beta_{j_g} \] (3.13)

\[ \sum_{g=1}^{2} (g) \beta_{j_g} = 0. \] (3.14)

The \( \beta_{j_g} \) may be estimated from the data using the \( n_{j_1j_2\ldots j_d}/n \), which are the maximum likelihood estimators of the parameters \( \pi_{j_1j_2\ldots j_d} = \pi_{j_1j_2\ldots j_d} / (1 - \pi_{22\ldots 2}) \) of the conditional likelihood based on the observed \( n \), the total number of elements observed by any of the \( d \) record systems. The resulting estimator of \( N \) is

\[ \hat{N}_d = \left( \prod_{g=1}^{d} n_g / n_{11\ldots 1} \right)^{1/(d-1)}, \ d \geq 2 \] (3.15)

where \( n_g \) is the total number of elements detected by the \( g \)th record system.

This natural generalization of the C-D (P-L) estimator was proposed by Chakraborty (1963). George (1968), using data for three sources compares the estimate given by (3.15) with the one which assumes a multinomial distribution given by Das Gupta (1964).

3.2.2 General Model Designs

The method of Section 3.2.1 is but one example of a more general procedure:

a) formulate a rank \( t < (2^d - 1) \) log-linear model for the \( \pi_{j_1j_2\ldots j_d} \)
b) estimate the parameters of the model from the unrestricted maximum likelihood estimators of $\hat{\pi}_{j_1 j_2 \ldots j_d}$ provided by the observed incomplete contingency table;

c) obtain estimates $\hat{\pi}_{j_1 j_2 \ldots j_d}$ of $\pi_{j_1 j_2 \ldots j_d}$ from the fitted model;

d) estimate $N$ by $\frac{n}{1-\pi_{22\ldots2}}$

These steps simply involve replacing the models of independence and quasi-independence in Section 3.2.1 with more general models which may incorporate types of source-dependence between subsets of the recording systems.

Explicitly, we consider the class of log-linear analysis of variance models for the conceptual $2^d$ contingency table, any member of which may be written as

$$\pi = \pi(\theta) = \exp (\theta X)/\sum_{\theta} \exp (\theta X)$$  \hspace{1cm} (3.16)

where $r = 2^d$,

$$\pi' = (\pi_{11\ldots1}, \pi_{11\ldots2}, \ldots, \pi_{22\ldots1}, \pi_{22\ldots2})$$;  \hspace{1cm} (3.17)

$\exp$ represents the elementwise exponential operator, $1^x$ is an $x$-vector of ones, $\theta$ is a vector of $t < r-2$ unknown parameters, and $X$ is a full rank $1 \times t$ design matrix of $1$'s and $-1$'s whose columns are jointly linearly independent of $1'$. The matrix $X$ is entirely analogous to the design matrix used to develop the fixed-effects analysis of variance of multi-way layouts within the context of regression analysis (the general linear model). The roles of factors in experimental design are taken by those variables (record systems here) which define the classifications described by each single dimension of the modeled contingency table; individual categories (or values, $j = 1, 2$) of each of these variables assume the role of factor levels, while log totals ($\log n_{j_1 j_2 \ldots j_d}$)
of individual cells are the basic observations. Thus, for example, the
design matrix \( X \) for complete independence among record systems is identical
to that for a main-effects model of a \( 2^d \) layout, with the column correspon-
ding to the general mean deleted; \( X \) matrices which include various types of
interaction in the general linear model may, with the general mean column
deleted, be used to introduce parameters representing corresponding types
of statistical source-dependence among record systems. Steps (b) to (d) may
be followed in order to obtain estimates of population size accounting for
such dependence in the observed data.

Fienberg (1972) and Bishop, Fienberg and Holland (1975) suggest the
use of matrices \( X \) representing members of the "hierarchical" class of models.
These models are characterized by the property that presence in the model of
an interaction parameter involving a subset \( S \) of the recording systems implies
the presence of additional main effect and interaction terms for every non-
empty subset of \( S \). Thus, if \( [X]_c \) is the set of columns in the matrix \( X \),
hierarchical models are exactly those for which

\[
[X]_c = \bigcup_{k=1}^{K} [X_k]_c \quad \text{for some } K, \quad \text{(3.18)}
\]

where \( X_k \) is the design matrix appropriate to a saturated factorial model
involving some subset \( S_k \) of the \( d \) record systems. These models correspond
to those ordinarily useful in the analysis of continuous experimental data,
where it is seldom appropriate to exclude an interaction term without also
excluding its higher-order "relatives." For example, with respect to the
triple-record system (TRS), the class of hierarchical models includes:

1) independence;

2) independence of one record system from the other two,
which are source-dependent;

3) conditional independence of two records systems, given
action of the third;
iv) pairwise source-dependence with no second-order interaction.

3.2.3 Stratification

If we try to minimize event correlation by partitioning the data into groups, \( P_i \) where \( i = 1, 2, \ldots, s \) by demographic characteristics, we obtain an incomplete \( 2^d \) table with one missing cell for each group, or an incomplete \( s \times 2^d \) table with \( s \) missing cells. The subscript \( i \) may be composite, representing combinations of values of different variables, so that this \( s \times 2^d \) table may be a re-expression of a higher dimensional cross-classification. The methods of Section 3.2.2 apply directly if we wish to model each subgroup and estimate the size of the corresponding population subgroup independently from the others. However, if we have some knowledge of the manner in which the demographic characteristics affect the operation of the various record systems, we may obtain more efficient estimates by incorporating this information into a joint model for all strata. If \( \pi_{1}^{*} \) is defined as \( \pi \) in Section 3.2.2 for stratum \( i \), and \( \pi_{1}^{*} = (\pi_{1}^{*}, \pi_{2}^{*}, \ldots, \pi_{s}^{*}) \), then we may write such a model as

\[
\pi_{1}^{*} = \exp \left( \beta_{1}^{*} \right) = D_{n}^{-1} \{ \exp \left( X_{1}^{*} \beta^{*} \right) \} .
\]  

(3.19)

In this formulation \( \beta^{*} \) is a \( u \times s(r-2) \) vector of unknown model parameters underlying the joint detection probabilities for all strata,

\[
X_{rs}^{*} = (X_{1}^{*}, X_{2}^{*}, \ldots, X_{s}^{*})
\]

(3.20)

is the composite design matrix, \( D_{n} \) is the diagonal matrix with the vector \( \chi \) as principal diagonal,

\[
\eta = [I_{s} \otimes I_{s}] \exp \left( X_{rs}^{*} \right) \]

(3.21)

and \( I_{s}, I_{s} \) and \( \otimes \) represent the \( v \)-identity, a \( v \times w \) matrix of units, and Kronecker multiplication, respectively. Analogously to Section 3.2.2, the composite design matrix \( X_{rs}^{*} \) is required to be of full rank, with columns
jointly linearly independent of the columns of $[1 \otimes I_s]$. Using this structure, we may incorporate into our model knowledge of the following types:

i. that a particular demographic factor alters the main effect parameters corresponding to some record systems but not others;

ii. that a particular demographic factor affects source correlations but not main effect parameters of certain pairs of record systems;

iii. that several demographic factors operate upon detection probabilities of recording systems in a particularly simple manner, i.e., no higher-order interactions between combinations of demographic variables and joint detection probabilities for several recording systems.

The use of assumptions of these types, where appropriate, results in estimates of total population size $N$ with lower variance than results from independent modeling of individual strata.

In addition, joint modeling of multiple strata provides a framework for application of another methodological tool, which we term "projection analysis." Log-linear model estimation of the total sample size $N$ is based on projections of the log-linear models for data collected in each stratum onto the missing cell of unobserved elements in that stratum. By projection analysis, we refer to the study of the effects of model choice, in terms of the selection of parameters within each individual stratum model and additional assumptions about across-stratum homogeneity, upon the set of projected counts of unobserved elements for the various strata. By varying the model used and studying the changes induced in the resulting projections, we may evaluate the sensitivity of ultimate estimates to the process of model selection. Further, through computer generation and analysis of data from a chosen model, the effects on stratum size estimates of alterations in recording systems which influence source correlations, or which differentially alter element detection rates in certain strata, can be simulated. The ideas behind this potentially useful tool for program planning appear in El-Khorazaty (1975).
3.3 Strategies for Fitting Log-Linear Models

In this section, we describe three strategies available for estimating parameters and fitted joint detection probabilities for the models of Section 3.2. Corresponding statistics for evaluating adequacy of fit are also referenced. In every case, once an acceptable model is decided upon, we use \( \hat{\pi}_{i;22\ldots2} \), the estimated probability that all record systems miss an element of stratum \( i \), to estimate the stratum size \( N_i \) through the formula

\[
\hat{N}_i = \frac{n_i}{1 - \hat{\pi}_{i;22\ldots2}}, \tag{3.22}
\]

where \( n_i \) is the total number of elements of stratum \( i \) recorded by any of the record systems.

The most general method involves a slight modification of the approach of Grizzle and Williams (1972) for fitting log-linear models to complete contingency tables, which they developed as an application of the general methodology described by Grizzle, Starmer and Koch (1969). Weighted least-squares (WLS) computational algorithms are applied to fit the postulated log-linear model to the observed vector \( \log \hat{p} \), where \( \hat{p} \) contains the various

\[
p_{i;j_1j_2\ldots j_d} = \frac{n_{i;j_1j_2\ldots j_d}}{n_i}, \tag{3.23}
\]

the unrestricted maximum likelihood estimates of

\[
\hat{\pi}_{i;j_1j_2\ldots j_d} = \frac{\pi_{i;j_1j_2\ldots j_d}}{1 - \hat{\pi}_{i;22\ldots2}}. \tag{3.24}
\]

The covariance matrix used is obtained by substituting the \( p_{i;j_1j_2\ldots j_d} \) for \( \pi_{i;j_1j_2\ldots j_d} \) in the asymptotic covariance matrix of \( \log \hat{p} \), determined by applying the \( \delta \)-method for deriving the covariances of transformed random variables (see Grizzle, Starmer and Koch (1969)). Thus, \( \log \hat{p} \) is expanded in a Taylor series about \( \log \bar{p} \), when \( \bar{p} \) is the corresponding vector of the
and the covariance matrix of the linear term extracted. This method yields a direct estimate \( \hat{\beta} \) of \( \beta \) without iteration; the fitted joint detection probabilities are obtained by substituting \( \hat{\beta} \) into the model equations. The estimator \( \hat{\beta} \) is a member of the class of procedures based on minimizing Neyman's (1949) modified chi-square criterion subject to a linearized hypothesis. As such, it is a Best Asymptotically Normal (BAN) estimate of \( \beta \). For moderate to large samples in practice, \( \hat{\beta} \) tends to be close to the estimate \( \hat{\beta} \) which maximizes the conditional likelihood based on the \( n_j \), in the sense that individual components of \( \hat{\theta} \) and \( \hat{\beta} \) tend to differ by less than the estimated standard deviation of either. The usual weighted least-squares algorithms produce test statistics, both for fit of the model and additional linear parametric restrictions, which belong to the class of test criteria defined by Wald (1943). Thus, in terms of asymptotic distribution and power, they are equivalent to the corresponding likelihood ratio tests based on the conditional likelihood. Finally, estimates of the covariance matrix of \( \hat{\theta} \), and that for the resulting fitted detection probabilities, may be obtained by further application of the \( \delta \)-method. All computations for the analysis just described may be executed using a general computer program for the analysis of categorical data, GENCAT (Landis, Stanish, Freeman and Koch (1976)), available from the University of Michigan.

When the postulated model is hierarchical, in terms of the entire set of dimensions involving both the different recording systems and the stratification variables, then the maximum likelihood estimators and likelihood ratio tests (based on the conditional likelihood) may be readily obtained.

In this situation

\[
[\hat{X}^*] = \bigcup_{m=1}^{M} [\hat{X}^*(m)]
\]

(3.25)
Where $X^*_{(m)}$ is the design matrix of a factorial model involving a subset of the dimensions determined by the recording systems and/or stratification variables. If $M$ is minimal, the observed marginal tables generating the $X^*_{(m)}$, $m = 1, 2, \ldots, M$, form a set of minimal sufficient statistics for the parameters of the model $X^*_{(m)}$ (Birch (1963), Bishop, Fienberg and Holland (1975)). The sufficient statistics not only generate the fitted table which maximizes the underlying likelihood, but in fact are reproduced by it; in other words, the maximum likelihood estimates of joint detection probabilities are the unique set of probabilities which both satisfy the model structure and generate marginal expected counts identical to the set of minimal sufficient statistics (Birch (1963)).

For some models, the fitted joint detection probabilities may be calculated explicitly and directly from the sufficient marginal tables. Generally, however, it is necessary to use a modification of the technique of "iterative proportional fitting" (IPF, or "raking") of Deming and Stephan (1940).

If $C_{(m)}$ is the observed marginal table generated by $X^*_{(m)}$, the technique is executed as follows:

i) form Table $T$, of the same dimensions as the observed incomplete table, with zeros in cells representing unobserved elements and units in all other cells.

ii) collapse $T$ to form the marginal array $C^{(1)}_{(1)}$ generated by $X^*_{(1)}$ from $T$; form $T^{(1)}_{(1)}$ by inflating each cell of $T$ by the ratio of its marginal category frequency in $C^{(1)}_{(1)}$ to that in $C^{(1)}_{(1)}$.

iii) for $m = 2, \ldots, M$, form $C^{(m)}_{(m)}$ from $T^{(m-1)}_{(m-1)}$ using $X^*_{(m)}$; form $T^{(m)}_{(m)}$ from $T^{(m-1)}_{(m-1)}$ using $C^{(m)}_{(m)}$.

iv) for $v \geq 2$, cycle through ii) to iii) substituting $T^{(v-1)}_{(m)}$
for $T, C^{(v)}_{(m)}$, for $C^{(1)}_{(m)}$, and $T^{(v)}_{(m)}$; continue until $T^{(v)}_{(m)}$
and $T^{(v-1)}_{(m)}$ are sufficiently close.

The elements of $T^{(v)}_{(m)}$ are then divided by the appropriate stratum sizes $n_i$
to yield the joint detection probabilities for each stratum. The estimated parameter vector $\hat{\beta}$ is obtained by substituting these into the model equations and solving, if desired. Estimated population sizes for each stratum, and likelihood ratio tests of fit associated with the model, or with comparisons of alternate models, may be calculated using the fitted probabilities, without ever explicitly obtaining the model parameters.

The estimates obtained by IPF may be preferable to those given by the WLS procedure when some observed stratum counts are modest, inasmuch as the asymptotic theory for the maximum likelihood estimators depends on the expected counts in cells of the marginal tables $C_{(m)}$, $m = 1, 2, \ldots, M$, where as that underlying the WLS approach depends on expected counts in individual cells of the incomplete table. All computations for the IPF-MLE analysis may be executed using a computer program for fitting log-linear models to contingency tables, ECTA, available from the University of Chicago.

When modest observed counts make the use of WLS unattractive and the proposed model is not hierarchical, estimates may often be derived by applying a Functional Asymptotic Regression Methodology (FARM) approach. This capitalizes on the observation that non-hierarchical models can be written as hierarchical models with linear restrictions on the parameters. Thus, we attempt to find an unsaturated hierarchical model from which the non-hierarchical model of interest may be derived through the imposition of linear restrictions. IPF is applied to derive an initial estimate of $\hat{\beta}$ whose sampling variability derives from the expected counts corresponding to margins of the observed data table rather than interior cells. WLS al-
gorithms are then applied to this preliminary estimate of $\beta$, using its estimated asymptotic covariance matrix under the hierarchical model (as determined by the $\delta$-method), to introduce the linear restrictions which reduce the hierarchical model to the more parsimonious non-hierarchical model of interest. The appropriate likelihood ratio test is used to assess fit of the initial hierarchical model, and a conditional WLS test used to evaluate adequacy of the subsequent reduction. Further reductions of the non-hierarchical model may be accomplished by the application of WLS to the fitted parameters.

The FARM procedure is somewhat simpler to implement computationally than to describe conceptually. Once an initial hierarchical model is chosen, fitted joint detection probabilities are obtained under this model by IPF. The FARM estimate of $\beta$ is then obtained by applying the WLS computational algorithms to the vector of these estimated probabilities instead of the observed proportion vector $p$. As a result, the FARM analysis may be performed by simple execution in sequence of the computer programs ECTA and GENCAT described previously.

The literature describing development of log-linear model theory and the fitting strategies outlined here is vast, and no attempt has been made in this review to document adequately the contributors. The application of log-linear model theory to the MRS-CMR problem is due to Fienberg (1972), and a full exposition of the IPF-MLE approach appears in Bishop, Fienberg and Holland (1975). The WLS approach was adapted and described by Koch, El-Khorazaty and Lewis (1977), while FARM procedures are due to Koch, Imrey, Freeman and Tolley (1977), and are applied to the MRS problem by El-Khorazaty, Imrey, Koch, and Lewis (1976).
4. The Capture-Mark-Recapture (CMR) Technique

4.1 Definition and Assumptions

The CMR technique has been used for more than three centuries for estimating the size of animal populations rather than human populations (Marks, Seltzer, and Krotki (1974)). It involves the selection (capture) at time $t_1$ of a sample of $n_1$ sampling units (individuals, animals, or events) from a population of unknown size $N$. Each sample unit is marked and returned to the population from which it was selected. Allowing enough time for marked units to be mixed with the whole population, a second sample of $n_2$ sampling units is drawn (recaptured) at time $t_2$, of which $n_{11}$ are marked. The estimated total population size, under the binomial assumption, is the same as the one given by (2.2). A third sample of $n_3$ units may be drawn at $t_3$ out of which $n_{111}$ units are recorded in the three samples and so on. In general, we have a multi-sample scheme where the estimate of the total population size, $N$, under the assumption of independence is given by (3.15) for $d$ samples. Both of these expressions are a consequence of the independence assumptions.

As mentioned in Section 2, the Peterson–Lincoln (P–L) index, developed for animal populations is the same as the (C–D) estimate proposed for human populations. Its approximate variance was obtained by Chandrasekar and Deming (1949). Later, Bailey (1951), using the binomial distribution, gave its exact variance, and Sen (1959) investigated the asymptotic convergence of the variance to the information limit. Otherwise, Robson and Regier (1964) stated that the bias of the P–L (C–D) estimator, which is of order $(1/n_{11})$, was essentially negligible if $n_{11} \geq 7$ and $(n_1 + n_2) < N$ since $(n_1 n_2 / N) > 4$.

Alternatively, Chapman (1951) proposed the estimator

$$\hat{N}_C = \frac{(n_1+1)(n_2+1)}{n_{11}+1} - 1$$

(4.1)
which is exactly unbiased when \( n_1 + n_2 \geq N \). Seber (1970) gave an estimator for the approximate variance of \( N_C \), and Wittes (1972) considered a variance estimator which is unbiased when \( n_1 + n_2 \geq N \). Two other modified estimates were suggested by Bailey (1951) for direct and inverse sampling. The robustness of some of these estimators has been studied by Seber (1970) and Holst (1971).

The estimates obtained from equation (2.2), or, more generally, equation (3.15) are based on a closed population for which it is assumed that no birth, death, or migration have occurred in the intervals between different samples, or, as Seber (1973) states, "that the experiment should be carried out over a short period of time, in fact, ideally at a single point in time." In addition, all events must have the same probability of being recorded in each sample, and marking must satisfy the matching assumptions described for human populations in Section 2. Further discussion of these assumptions in the CMR context can be found in DeLury (1954), Scattergood (1954), Cormack (1968), Carothers (1971), and Seber (1973). In many situations, these assumptions cannot be tested (Darroch (1959), Seber (1965), and Cormack (1968)). Thus, attention is directed at the magnitude of the bias as undertaken by Cormack (1972) for the Seber-Jolly (S-J) models developed by Seber (1965) and Jolly (1965) for open animal populations. Otherwise, Cormack (1966) suggests a test for equal catchability, and Holst (1972) and Mohamed (1973) obtain estimates of the population size in a framework where an event which had never been caught before and an event with previous history of capture may have different probabilities of capture in the respective samples.

Another disadvantage of the P-L (C-D) index, especially with respect to DRS applications, is the assumption that sample sizes are pre-determined (Kindahl (1962)). This assumption is not true either because of the nature of the registration systems and/or the use of cluster sampling so that the sample sizes are random variables. The relaxation of this assumption requires the inclu-
sion of the probabilities of obtaining two samples of size \( n_1 \) and \( n_2 \) from the total population size \( N \) in the probability distribution functions (see El-Khorazaty, 1975).

4.2 Mathematical Modeling

Different CMR models have been developed for closed and open animal populations. For closed populations, in which the only random or stochastic element is the sampling procedure in use, Cormack (1968) states that no study will give statistically valid estimates unless the sampling procedure is specifically aimed at the particular species under study and the behavior of that species is understood. Under the assumptions of simple random sampling, no loss of marks and independence of the two samples, the conditional probability distribution of \( n_{11} \) jointly recorded events from \( n_1 \) and \( n_2 \) events for the respective sources is the hypergeometric distribution

\[
P[(n_{11}, n_1) | (n_2, N)] = \binom{n_1}{n_{11}} \binom{N-n_1}{n_2-n_{11}} / \binom{N}{n_2},
\]

\( n_{11} = 0, 1, \ldots, \min(n_1, n_2) \).

The model (4.2) was discussed by Chapman (1951) and Sen (1959). However, it assumes sampling without replacement, which does not hold for the MRS technique as it is used in demographic and epidemiologic research where events are recorded by different sources but not captured. In this regard, Seber (1973) stated that the model assuming sampling with replacement "would apply when the animals [events] are merely observed and not actually captured," which is the case in human populations. Therefore, Bailey (1951) used the following binomial approximation to the hypergeometric distribution (4.2):

\[
P[(n_1, n_2, n_{11}) | (N, p_1, p_2)] = \binom{n_2}{n_{11}} \binom{n_1}{n_1-n_{11}} (1-p_{11}) \binom{n_2-n_{11}}{n_2-n_{11}}
\]

\( 0 < p_1, p_2 < 1; \ n_{11} = 0, 1, \ldots, \min(n_1, n_2) \),

where \( p \) is the probability that an event has been recorded by the \( g \)-th source.

For the two distributions (4.2) and (4.3), the maximum likelihood estimate of
N is the P-L (C-D) estimate given by (2.2). However, because of its bias, Bailey (1951) and Chapman (1951) proposed other modified estimates, as mentioned in Section 4.1.

The probability distributions (4.2) and (4.3) may be extended directly to the general case of d samples (recording sources). Chapman (1952), under the same assumptions previously indicated for the two-sample case, obtained the joint probability function as the product of hypergeometric distributions. Using the binomial distribution, Darroch (1958) obtained the maximum likelihood estimate for N for the multiple-recapture census case. Berg (1974, 1975, 1976) used the ratio estimate proposed by Pathak (1964) to derive a uniformly minimum variance unbiased estimator for this case. Seber (1962) studied the multi-sample single recapture census situation, where releases and recaptures are made at different times, and Manly (1974, 1975) derived maximum likelihood estimates of survival probabilities. Darroch (1961) discussed the stratified sample case. Finally, incomplete contingency table methods for the multiple-recapture census were reviewed in Section 3.

Open (dynamic) population models should be considered when sampling continues over an appreciable period of time. This issue may be seen, for the case of two sources, by defining $\phi_{12}$ as the probability that an event captured in the first sample will be alive and in the population at time $t_2$. Then,

$$E \left\{ \frac{(\phi_{12} n_1) n_2}{n_{11}} \right\} = \phi_{12} N$$

or

$$E(N) = N.$$  \hspace{1cm} (4.4)

This formula means that as long as $\phi_{12}$ is the same for marked and unmarked events, the P-L (C-D) estimate is still applicable. However, $n_{11}$ will be smaller than expected if there are deaths or emigrants among the $n_1$ events; hence, $\hat{N}$ will tend to underestimate the population size. Alternatively, the effect of the new events $B_{12}$ (birth or immigration) that joined the popula-
tion in the interval \((t_1, t_2)\) can be seen from the following relation:

\[
E\left( \frac{\phi_{12} n_1 n_2}{n_{11}} \right) = \phi_{12} N + B_{12}
\]  \hspace{1cm} (4.5)

or

\[
\hat{E}(N) = N + B_{12}/\phi_{12},
\]

which shows that the P-L (C-D) estimate of \(N\) tends to over-estimate the initial population size. Therefore, deterministic and stochastic models have been introduced for open populations. Many authors have dealt with the case of deterministic models, for example, Jackson (1937, 1939, 1948), Leslie and Chitty (1951, 1952), and Leslie, Chitty, and Chitty (1953). The most general deterministic model for multiple recapture data was given by Jolly (1963). This model allows for both "death and dilution" and assumes a multinomial distribution conditional on the previous results.

Stochastic CMR models were first introduced by Hammersley (1953). However, as shown by Darroch (1958), the likelihood function used by Hammersley was incorrect and gave the estimate of the population size at any time as the number of events captured up to that time. Hence, Darroch (1959) obtained maximum likelihood estimates for the case where either death (emigration) or immigration (birth), but not both, were affecting the population. These models have been shown to be special cases of the general models suggested by Jolly (1965) and Seber (1965). Both of these models give identical estimates for the population size, the number of marked events, the number of new immigrating events, the probability of leaving the population, and the probability of being captured. The variances of these estimates were derived. The S-J model is the only one which uses all the information provided by the experiment. These two stochastic models are chosen over deterministic models developed earlier since the latter are just approximations to real life by requiring certain entities to be integers, provide less valid results, and provide no estimate of the variability of the estimates. The results ob-
tained using the S–J model are based on the assumption that each event has equal catchability and the same probability of survival regardless of previous capture history. The relaxation of the assumption of equal catchability was considered by Holst (1972) and Mohamed (1973), while the weakening of the assumption of equal survival rates was considered by Robson (1969). A recent paper by Pollock (1975) dealt with a further generalization which allowed unequal catchability and unequal survival. The S–J and Robson models are special cases of Pollock's general theory.

More specifically, if the number of events recorded is assumed to be the same as the number released, Jolly's model for d samples is proportional to

\[
\prod_{g=1}^{d} p_g^{N-g} \left(1-p_g\right)^g = \phi_g^{g+1_{g+1}} \left(1-\phi_{g+1}^{g+1}\right) + B_{g+1}^{g+1} \quad (4.6)
\]

in which \(N\) is the total number of events in the population alive just before time \(t\). Other relevant assumptions have been mentioned previously. This distribution was obtained by Seber (1965) using the multinomial rather than the binomial distribution.

Adopting Jolly's approach, El–Khorazaty and Sen (1976) developed the following conditional probability distribution for estimating the total number of events in a DRS as the product of five binomial distributions

\[
P[(n_1, n_2, n_{11}, n_{21}, n_{11}) | (N_1, p_1, p_2, \phi_{12}, B_{12})] = \left\{ \begin{array}{c}
\prod_{g=1}^{2} p_g^{N-g} \left(1-p_g\right)^g \\
\left(\frac{n_1}{N_1}\right)^{n_{11}} \left(1-\frac{n_1}{N_1}\right)^{n_{21}-n_{11}} \\
\left(\frac{n_{11}}{N_{11}}\right)^{n_{11}} \left(1-\frac{n_{11}}{N_{11}}\right)^{n_{21}-n_{11}} \\
\left(\frac{n_{21}}{N_{21}}\right)^{n_{21}} \left(1-\frac{n_{21}}{N_{21}}\right)^{n_{21}-n_{11}} \\
\left(\frac{n_{21}}{N_{21}}\right)^{n_{21}} \left(1-\frac{n_{21}}{N_{21}}\right)^{n_{21}-n_{11}} \end{array} \right\} .
\]

(4.7)

where

\(N_{11}\) is the total number of events recorded at \(t_1\) and alive at \(t_2\), and

\(N_{21}\) is the total number of events in the population at \(t_2\) and unrecorded at \(t_1\). We define \(N_{21} = N_{21} + N_{11}\).
This model has the following characteristics:

1) the two sources are independent;
2) \( n_1 \) and \( n_2 \) are not pre-determined;
3) sampling is with replacement;
4) the population of events is homogeneous;
5) the number of new events joining the population in the interval \( (t_1, t_2) \) is \( B_{12} \);
6) the matching probability is modeled.

Seber's approach treated certain parameters as unknown and fixed rather than as random variables like Jolly's. Moreover, \( N_g \) and \( B_{12} \) are considered as functions of the other estimates. Assuming the multinomial distribution, El-Khorazaty and Sen (1976) developed the probability distribution for this situation as

\[
P[(n_{12}, n_{21}, n_{11}) \mid (N_1, N_2, p_1, p_2, \phi_{12})] \\
= k_1 \left\{ \frac{2 \pi}{\prod_{g=1}^{N_2} (1-p_g)^{N_g-n_g}} \right\} \left\{ \phi_{12} \left[ \frac{1 - \phi_{12} p_2}{1 - p_2} \right] \right\}^{n_{12}},
\]

where

\[
k_1 = \frac{N_1! (N_2-n_1)!}{(N_1-n_1)! [N_2 - (n_{12} + n_{21} + n_{11})]! n_{12}! n_{21}! n_{11}!},
\]

and \( N_2 - N_1 \) is the net increase in the unrecorded population yielding \( N_{21} = N_1 n_1 + N_2 - N_1 = N_2 - n_1 \). This model is equivalent to model (4.7) except that it does not include:

i) the matching probability, since Seber's approach deals with \( n_{12}, n_{21}, \) and \( n_{11} \) rather than \( n_1 \) and \( n_2 \), and

ii) \( B_{12} \), since it is a function of other estimates.

Finally, for the case of two correlated samples (sources), El-Khorazaty and Sen (1976), following the approach presented by Seber (1970), developed the following probability distribution:
\[ P[(n_{12}, n_{21}, n_{11})| (N_1, N_2, p_{12}, p_{21}, \phi_{12})] = k_1 \left\{ \left( \frac{n_{11}}{p_{11}} \right)^{n_{11}} \left( \frac{p_{11}\phi_{12}}{p_{1}} \right)^{n_{12}} \right\} \left\{ \left( \frac{n_{21}}{p_{21}} \right)^{n_{21}} \left( \frac{N_2-n_{21}+n_{11}}{1-p_{21}} \right)^{N_2-(n_{11}+n_{21}+n_{11})} \right\} \] (4.9)

where \( k_1 \) is the same as in (4.8), \( p_{12} \) is the number of events recorded in the first but not in the second sample, and \( p_{21} \) is vice versa. Similar models were developed by El-Khorazaty and Sen (1976) for the triple-record system (TRS) under either the assumption of independence or dependence among the different sources.

### 4.3 Epidemiologic and Demographic Applications

Applications of the CMR technique to epidemiological data may be found in Wittes and Sidel (1968), Lewis and Hassanein (1969), Wittes (1970, 1974), and Wittes, Colton, and Sidel (1974). Examples given deal with (1) the surveillance of infectious disease reported by physicians, nurses, and bacteriology laboratories, (2) the estimation of the total number of patients receiving methicillin (a synthetic penicillin) as reported by nurses, medication sheets, and pharmacists, and (3) infants with a specific congenital anomaly reported by obstetric clinics, hospitals, departments of public and mental health, and public and private schools. These different applications showed the CMR methodology to be helpful in many ways. Using available, but incomplete, records involves minor costs when compared with other alternatives, such as sampling from the general population, which can often be prohibitively expensive. Also, these applications were informative with respect to the relative effectiveness of different recording systems in identifying events "characterized by a trait which occurs only rarely in the population at large" (Wittes (1974)). Measuring relative effectiveness was very important "to establish a means to evaluate the monitoring system itself" (Lewis and Hassanein (1969)) and to improve the quality of medical care by arriving at more accurate estimates of prevalence of a specific disease.
In demographic fields, the CMR strategy was used as a method for estimating the number of events (e.g., birth and death) in a human population by El-Khorazaty (1975) and El-Khorazaty and Sen (1976). The CMR stochastic models, developed by Seber and Jolly, have been adjusted and generalized to human populations. Models were developed as shown in Section 4.2, for the systems (DRS and TRS, respectively) for situations involving dependence among the different sources, which is analogous to unequal catchability (due to mark selectivity) in animal populations. Demographic applications of the developed likelihood functions have adjusted for the existing dependence without prior knowledge about its magnitude, direction, or underlying causes, and thus provide an effective framework for bypassing the under- or over-estimation caused by positive or negative correlation among the sources of information.

A wide variety of applications of the CMR strategy may be available whenever there is a set of events in a specific environment that are not self-reported, in which case, estimates for their total number become of interest. Other applications would be occurrences of accidents or crimes, diagnoses of rare diseases, existence of jobs or programs, etc.

5. **Comparing the MRS (DRS) and CMR Techniques**

The two techniques are related since both are methods of adjustment of the available data. While the purpose of the CMR is estimation of population size, the estimation of the total number of events is usually of interest in the MRS. Actually this is not a difference as such since it depends on the definition of the population under consideration, i.e., whether it is the whole population or the population of events.

Some writers believe that there is a wide divergence between the two techniques in the matching process (Marks, Seltzer and Krotki [1974]). This is because the MRS is concerned essentially with the event recorded by different sources, and hence with a two-way match (which determines the matching status of all the events in both sources), while the CMR is concerned with determining
the group to which an animal belongs (captured or uncaptured), and hence with a one-way match. From the methodological point of view, this is true but practically the two techniques lead to the same result since in the MRS we are looking for the numbers recorded by different sources (using a specific matching rule) in spite of the use of events' characteristics to get these counts. Thus, our interest in a specific event is a means to an end and not a goal in itself.

Similarities and dissimilarities between the CMR and MRS techniques arise mainly because of their application to different types of populations. Instead of marking in animal populations, the characteristics (e.g., name, sex, date of occurrence, address, etc.) of each event are recorded in the case of human populations. Missing events have more importance in the theory of the CMR for animal populations than in the MRS for human populations because less than half of the population is usually captured in the CMR, while, as Seltzer (1969) shows, match rates usually exceed 85 percent of the true number of events in the DRS. Thus, the two techniques address related problems: (a) in the CMR there are two or more incomplete censuses of the animal population of an area, and (b) in the MRS there are two or more methods (surveys) of incomplete recording of the events in an area. Moreover, Table III indicates the relationship between the assumptions mentioned in the literature for animal populations (CMR technique) and the equivalent assumptions for human populations (MRS technique).
### Table III. Comparison Between The Assumptions Required For Human and Animal Populations

<table>
<thead>
<tr>
<th>Human Populations</th>
<th>Animal Populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No geographic or time coverage bias in recording system.</td>
<td>Animals have the same chance of being recaptured whether marked or not.</td>
</tr>
<tr>
<td>2. No recording error, due to respondent or interviewer, for events occurred.</td>
<td>Animals do not lose their marks.</td>
</tr>
<tr>
<td>3. Perfect matching criterion which does not allow for erroneous matches or erroneous non-matches.</td>
<td>Sampled animals are classified correctly as marked or unmarked.</td>
</tr>
<tr>
<td>4. No response correlation bias, i.e., failure to maintain independence between the two sources of information.</td>
<td>Every animal, marked or unmarked, has the same chance of recapture, i.e., marked animals are randomly mixed with the unmarked ones and marking does not affect catchability.</td>
</tr>
<tr>
<td>5. No events (birth, death or migration) occur during the sampling period, i.e., no new events are added to the population between the first and second sampling.</td>
<td>The proportion of marked animals in the population remains constant during the sampling period, i.e., no marking is done during this period and the population is closed.</td>
</tr>
<tr>
<td>6. Death rates are the same for events recorded or not recorded.</td>
<td>Allowance for mortality should be conditioned so that it is the same among marked and unmarked animals.</td>
</tr>
<tr>
<td>7. The proportion of individuals having the event remains constant during the sampling period. Or, in case of estimation for the whole population, the vital rates are the same in the population as in the sample.</td>
<td>The proportion of marked animals in a sample is equal to the proportion of marked ones in the population.</td>
</tr>
</tbody>
</table>
In view of the similarities of the assumptions for the two techniques, the possibility emerges for the use of the theory already developed for the CMR technique for estimating the total number of events for human populations. Choi (1970), and El-Khorazaty and Sen (1976) employ this theory to modify the C-D estimates, developing several likelihood functions for the DRS and the TRS techniques as shown in Section 4.2. While El-Khorazaty (1975) and El-Khorazaty and Sen (1976) allow the occurrence of birth (immigration) and death (emigration) and assume dependence among the different sources to arrive at their modifications, Choi (1970) uses the Bayesian estimation procedure for this purpose. In addition, the coverage of more than 85 percent of the true number of events by the MRS technique will give more statistically precise estimates with smaller variance (Cormack [1972]). This is because, as DeLury (1954) states, "the success of these methods [CMR methods] depends on capturing both a considerable number and a considerable proportion of the individuals in a population." Choi (1970) shows that his estimates have smaller mean square errors than the C-D estimates.

A final methodological difference between the MRS and the CMR techniques lies in the sampling procedure applied. While the MRS is concerned either with the whole population or usually with an area sampling, the CMR is usually based on simple random sampling.

Acknowledgements

This research was in part supported by the U.S. Bureau of the Census through Joint Statistical Agreement JSA 76-93 and the Population Council, New York. The authors would like to thank Jean McKinney for her conscientious typing of this manuscript.
References


Wald, A. (1943). Tests of statistical hypotheses concerning general parameters when the number of observations is large. Transactions of the American Mathematical Society, 54, 426-482.


Yesilcay, Y. (1975). The mean square error of the estimate as a criterion for the assessment of alternative approaches to matching in a dual record system. University of North Carolina Institute of Statistics Mimeo Series No. 1035, Chapel Hill, N.C.