ISSN: 2641-6336
DOI: 10.33552/ABBA.2022.04.000599
Annals of
Biostatistics & Biometric Applications
Iris Publishers

Research article

Effect of Errors in Dual Record System Estimates

H Oztas Ayhan*

Professor Emeritus, Department of Statistics, Middle East Technical University, Turkey

***Corresponding author:** H. Oztas Ayhan, Professor Emeritus, Department of Statistics, Middle East Technical University, Turkey.

Received Date: November 3, 2022

Published Date: November 22, 2022

Abstract

This article examines the direction and magnitude of biases involved in the application of dual record systems for estimating vital events. During this study, three sources of error are considered; inclusion of spurious reports of vital events, matching errors, and lack of statistical independence of the two data collection systems. Since the biases produced by each of these components may be of opposite signs, the total bias may be less than the error produced by a single source. Expressions for the relative bias from each of the components are developed and combined in a model for total relative bias.

Keywords: Correlation bias; Dual record systems; DRS estimators; Matching errors; Measurement errors;Registration systems; Sample surveys; Spurious reports; Statistical independence; Vital statistics

Introduction

Dual record system (DRS) estimator was first proposed by Chandra Sekar and Deming (CD) in 1949 for estimating vital events in demographic surveys. For many years, it has been widely used in many developing countries, where the vital registration systems were not established widely. Dual record system estimation was used during the 1960s and early 1970s as an alternative method of data collection to single round surveys in developing countries due to their coverage problems and lack of data quality. Further methodological studies on the dual record systems estimators has led statisticians and demographers to use other survey techniques in place, because of the weakness of the statistical assumptions of the model as well as the biases involved in estimators. In this study, a review of the available literature was made and three sources of error in DRS are examined. These are inclusion of spurious reports of vital events, matching errors, and lack of statistical independence of the two data collection systems.

Dual Record System Estimation

The CD method is based on collecting and matching data for the same units of observation, from two independent data collection sources. These two data sources may be any combination of the *registration, census* or *sample surveys*. Collecting vital information from two independent sources and matching these records by a 2 x 2 contingency table provides the layout in Table 1.

Table 1: The Data Collection Structure of Dual Record System.

		Data Source 2		
	Type of Report	**Reported**	**Not Reported**	**Total**
Data Source 1	Reported	n_{11}	n_{12}	n_{1+}
	Not Reported	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

The Conventional Estimators

The classical Chandra Sekar - Deming Estimator,

$$\hat{n}^{(CD)} = n_{11} + n_{12} + n_{21} + \hat{n}_{22}^{(CD)} \quad (1)$$

where $\hat{n}_{22}^{(CD)} = n_{12}n_{21} / n_{11}$ (2)

On the other hand, same estimator can be presented in the following form by using the information on marginal totals,

$$\hat{n}^{(D)} = (n_1 + n_2)(n_1 + n_2)/n_1 = n_{1+}n_{+1}/n_1 \quad (3)$$

In the past years, excellent works are also done on the dual record system assumptions, matching errors, estimators, sampling variances and the biases by [1-5] had examined the problems related to the CD estimators and had proposed some improvements.

Assumption of Independence

[6] proposed their model on the assumption that the chances of an event being missed by the two data sources are independent of one another. In other words, the basic Chandrasekar-Deming model explicitly assumes that the conditional probability of an event being caught by one system, given that it is caught by the second system, is equal to the conditional probability of an event being caught by the first system given that it is missed by the second system. In practice, it is unlikely that these conditional probabilities are really equal [2].

When the total number of events which are missed by both data collection systems are determined on the basis of $\hat{n}_{22} = n_{12}n_{21} / n_{11}$ then it consequently implies independence. This also means that, the number of vital events which are missed by both systems will be the same as one another, in amount, which does not seem realistic.

Chandra Sekar and Deming observed that it may be possible to reduce the bias resulting from lack of independence by classifying the events into homogenous groups on the basis of age, sex and other appropriate characteristics and making the estimate of events separately for each group. This will be effective if the correlation for the contingency table for each grouping or stratum is near zero but the correlation for the contingency table for all strata combined is not zero [7].

Measures of Correlation

Chandra Sekar and Deming had recognized that the assumption of zero correlation was unrealistic and recommended the procedure of calculating separate estimates for events classified in homogeneous groups in order to reduce the correlation. [7] made some comments on the results from many available studies that, it was not clear to what extent this grouping procedure was used in preparing estimates. These limited results indicated the likely existence of a positive correlation for the overall population of events. Even if one could assume that all efforts were successful in having the two data collections carried out by different people, with neither group having any access to the results obtained by the other group, one would still not have achieved a situation with zero correlation or statistical independence. The latter requires, as stated by [6], that the events missed by either collection be equivalent to a simple random sample of all events occurring in the survey population. On a priori grounds it is reasonable to believe that for certain classes, the events would be more difficult to observe or record than for others, regardless of the method used.

in order to clarify their point, they assume that both collections use the same procedure and as a result the expected value of n_{12} and n_{21} are the same, while for the sample $n_{12} \cong n_{21} = n^*$. These are the number of not observed vital events by one source which are observed by the other source.

Jabine and Bershad (1968) proposed the following estimator.

$$\hat{n}^{(B)} = n_1 + 2n^* + \frac{(n^*)^2}{n_1} = \frac{(n_1 + n^*)^2}{n_1} \quad (5)$$

The correlation between the two sets of observations was determined as,

$$\hat{\rho}^{(B)} = \frac{n_1 n_2 - (n^*)^2}{(n_1 + n^*)(n_2 + n^*)} \quad (6)$$

This expression also indicated that the CD estimate of $\hat{n}_{22}^{(CD)}$ which is $(n^*)^2 / n_{11}$ will be an underestimate if the correlation is positive [7].

On the other hand, [8] evaluated the developments of the recent past in their paper. They emphasized that, $\hat{n}_{22} \neq n_{12}n_{21} / n_{11}$ if the correlation is present. The correlation coefficient of n_{22} is defined as;

$$\rho_{n_{22}}^{(D)} = \frac{n_{11}n_{22} - n_{12}n_{21}}{[(n_{11} + n_{21})(n_{12} + n_{22})(n_{11} + n_{12})(n_{21} + n_{22})]^{1/2}} \quad (8)$$

They proposed that, the CD estimator

$\hat{n}_{22}^{(CD)} = n_{12}n_{21} / n_{11}$ if $\rho_{n_{22}} = 0$.

If $\rho_{n_{22}} > 0$, then $\hat{n}_{22}^{(CD)} > n_{12}n_{21} / n_{11}$ and the CD method underestimates $\hat{n}_{22}$.

If $\rho_{n_{22}} < 0$, then $\hat{n}_{22}^{(CD)} < n_{12}n_{21} / n_{11}$ and the CD method overestimates $\hat{n}_{22}$.

Main Sources of Bias

There are several sources of bias related to DRS estimation. [9] pointed out the main sources as: correlation between the two sources, coverage error, matching error, and denominator error. Most of the difficulties with the method as currently practiced arise from the need to ensure identical coverage by the two sources without loss of independence. All four sources of bias are involved in this, and the heavy investment in mapping besides the elaborate matching operation arise in efforts to combat this problem. In practice, one often finds that any particular proposal for improving the coverage fit of the two sources raises the spectra of "loss of independence". [1] also classified the sources of bias related to DRS estimation in a similar manner. [2] and [1] also studied the *spurious reports, matching errors, lack of independence* and *interaction of sources of error* in the CD estimate. They presented formulations for the *relative bias, limits of bias due to the inclusion of spurious reports,* and *net match errors.* [7] proposed the following expression for the *bias* of the CD estimate,

$$B(\hat{n}^{(B)}) = \hat{n}^{(D)} - (n_1 + 2n^* + n_2)$$

$$= \frac{(n^*)^2}{n_{11}} - n_{22} \qquad (9)$$

and the *relative bias* is estimated by,

$$B^{(B)} = \frac{\frac{(n^*)^2}{n_1} - n_2}{n_1 + 2n^* + n_2} = \frac{-\hat{\rho}}{1-\hat{\rho}} \frac{n^*}{n_1} \qquad (10)$$

[7] provides information in Table 2 to show the factors which affect the *bias* of the CD estimate.

Source: [7]

Table 2: Relative Bias of CD Estimate for Selected Values of ρ and [$n_1 / (n^* + n_1)$].

ρ	$n_1 / (n^* + n_1)$				
	0.5	**0.6**	**0.7**	**0.8**	**0.9**
0	0	0	0	0	0
0.05	0.053	0.035	0.023	0.013	0.006
0.1	0.111	0.074	0.048	0.028	0.012
0.15	0.176	0.118	0.076	0.044	0.02
0.2	0.25	0.167	0.107	0.063	0.028
0.3	0.429	0.286	0.184	0.107	0.048
0.4	0.667	0.444	0.286	0.167	0.074

Source: Jabine and Bershad (1968)

Matching Bias

Matching bias is the result of imperfect matching of persons in the household survey and the census. In matching the events identified by the two procedures, we may err in either direction. That is, we may fail to detect a match, or we may declare a match when none exists. [7] stated that, if the net effect of matching errors is to understate n_1 , then we will overestimate the true number of events in the same proportion. In the presence of a positive correlation we have,

$$\hat{n}^{(B)} = N \frac{n_1}{n_1'}\left(1 - \frac{\rho}{1-\rho}\frac{n^*}{n_1}\right)$$

If the net effect of matching errors is to overstate n_1 , that is $n_1' > n_1$, then $\hat{n}$ will always be an underestimate. In the case of that $n_1' < n_1$, the bias resulting from matching errors will act in the opposite direction from the bias due to correlation, so that it is not possible to generalize about the joint effect of the two kinds of bias [7].

[2] reports that, the *relative bias* in the Chandra Sekar – Deming estimate due to *net match error* may be expressed by;

$$B(\hat{n}^{(D)}) = -\frac{m}{1+m}$$

where B : relative bias of CD estimate ($\hat{N}$)

m : net relative matching error $\left[\left(n_1 / n_1^{(T)}\right) - 1\right]$

If *m* is positive, on balance there are more false matches than false non-matches, then $\hat{N}$ is biased downward. If *m* is negative, there are more false non-matches than false matches, then $\hat{N}$ is biased upward [2].

Out of Scope Bias

Out of scope bias refers to erroneous inclusions of persons by either the survey or the census. Some events may be recorded by either or both procedures when in fact they are not included in *N*, the true number of events associated with the units of observation. When recorded by only one method, such erroneously included

events will always increase the size of $\hat{n}^{(D)}$, and hence tend to overstate the true number of events. [7] proposed that, when recorded by both methods, *erroneously included events* may affect the size of $\hat{n}^{(D)}$ depending on whether.

$n_1 (n_1 + 1) > (n^*)^2$ increases the size of $\hat{n}^{(D)}$

$n_1 (n_1 + 1) < (n^*)^2$ decreases the size of $\hat{n}^{(D)}$

$n_1 (n_1 + 1) = (n^*)^2$ the size of $\hat{n}^{(D)}$ is unchanged.

Spurious Reports

The model proposed by Chandra Sekar and Deming assumes that only reports of proper events, as defined for a given study, are included when making the estimates. In practice, this is not the case. One or both systems may stray beyond the geographical, temporal, or conceptual boundary being employed. These kinds of errors result in spurious reports which in turn affect the estimate of total events.

[2] reports that, considered independently from other sources of error, spurious reports in either or both systems wil bias Chandra Sekar – Deming estimates of the number of vital events in an upward direction. Assuming the statistical independence of the two systems and that no matching errors occur, the *relative size of CD bias* is;

$$B(\hat{n}^{(D)}) = [U / V_1 V_2] - 1$$

where U: matches based on valid reports as a proportion of all matches,

V_1 : valid reports of system 1 as a proportion of all reports of system 1

V_2 : valid reports of system 2 as a proportion of all reports.

If all matched reports truly represent within-scope events, then the *relative bias*

of CD simplifies to.

$$B(\hat{n}^{(D)}) = [1 / V_1 V_2] - 1$$

The size of the bias due to the inclusion of spurious reports in the Chandra Sekar – Deming estimate of the vital events lies within the following *limits*.

$$[(1 / V_2) - 1] \leq B(\hat{n}^{(D)}) \leq [(1 / V_2^2) - 1]$$

Coverage and Nonresponse Bias of Estimators

The following estimators are proposed by [10,11] for the case of equal sample sizes ($n_1 = n_2$).

Estimators which are based on data source 1 will be,

$\hat{n}^{(A)} = n_{+1} + n_2 + [n_1^* - n_{1+}) - n_2]$ where $[n_1^* - n_{1+}) - n_2] = \hat{n}_2^{(A)}$

For data source 1, the *amount of bias* due to coverage and nonresponse errors will take the following form,

$$B(\hat{n}^{(D)}) = [n_1^* - n_{1+}) - n_2] - \hat{n}_2^{(D)}$$

The *relative bias* of the estimator for this case will take the following form,

$$B(\hat{n}^{(D)}) = \hat{n}_2^{(D)} - [n_1^* - n_{1+}) - n_2] / \hat{n}_2^{(D)}$$

Similar *estimators* can also be determined on the basis of data source 2 which will be,

$\hat{n}^{(A)} = n_{1+} + n_2 + [n_2^* - n_{+1}) - n_2]$ where $[n_2^* - n_{+1}) - n_2] = \hat{n}_2^{(A)}$

For data source 2, the amount of *bias* due to coverage and nonresponse errors will take the following form,

$$B(\hat{n}^{(D)}) = [n_2^* - n_{+1}) - n_2] - \hat{n}_2^{(D)}$$

The *relative bias* of the estimator for this case will take the following form,

$$B(\hat{n}^{(D)}) = \hat{n}_2^{(D)} - [n_2^* - n_{+1}) - n_2] / \hat{n}_2^{(D)}$$

By using n_1^* or n_2^* as the base, [10] had proposed the following estimator of the last cell of the design, which has *coverage error* and *nonresponse error* components.

$$\hat{n}_2^{(A)} = \{[N(n_1) \wedge N(n_2)] + [N(n_1) \wedge U(n_2^*)] + [U(n_1^*) \wedge N(n_2)] + [U(n_1^*) \wedge U(n_2^*)]\}$$

Conclusion

The proposed estimators for the dual record systems are based on further division of the cells of the original table and the results have shown that, they improve the underestimation of the total counts when compared with the classical Chandra Sekar-Deming Estimator. When a reasonably accurate registration system is available in a country, then the dual record system approach will naturally be less costly than a high quality based single round survey. In addition to their use for determining the vital events, the DRS estimates can also be used for measuring the population census coverage errors as well as to improve the total counts in other surveys.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.

References

1. Marks E S, W Seltzer, K J Krotki (1974) Population Growth Estimation: A Handbook of Vital Statistics Measurement. The Population Council, NY.
2. Seltzer W, A Adlakha (1974) On the Effect of Errors in the Application of the Chandrasekar-Deming Technique. Laboratories for Population Statistics, Reprint Series 14: 13.
3. Seltzer W (1969) Some Results from Asian Population Studies. Population Studies 23: 395-406.
4. Greenfield C C (1975) On the Estimation of a Missing Cell in a 2 x 2 Contingency Table. Jour. Royal Statist. Society A 138(1): 51-61.
5. Greenfield C C, S M Tam (1976) A Simple Approximation for the Upper Limit to the Value of a Missing Cell in a 2 x 2 Contingency Table. Jour. Royal Statist. Society A 139(1): 96-103.
6. Chandra Sekar C, W E Deming (1949) On a Method of Estimating of Birth and Death Rates and the Extent of Registration. Jour. Amer. Statist. Assoc 44: 101-115.

7. Jabine T B, M A Bershad (1968) Some Comments on the Chandrasekaran-Deming Technique for the Measurement of Population Change. Proceedings of the CENTO Symposium on Demographic Statistics p. 189-206. Karachi, Pakistan.

8. Chandrasekaran C, W E Deming (1981) On the Correlation Bias in the Application of Chandra-Deming Method for Estimating Vital Events. Cairo Demographic Centre, Working Paper 2: 31.

9. Scott C (1974) The Dual Record (PGE) System for Vital Rate Measurement: Some Suggestions for Future Development. In International Population Conference, Liege 1973, Volume 2. International Union for the Scientific Study of Population. p. 407-416.

10. Ayhan HO (2000) Estimators of Vital Events in Dual Record Systems. Journal of Applied Statistics 27(2):157-169.

11. Ayhan HO (1997) Alternative Estimators for Dual Record Systems. Bulletin of the International Statistical Institute 57(2): 305-306.

12. Blacker J G C (1977) Dual Record Demographic Surveys: A Reassessment. Population Studies 31(3): 585-597.

13. Cleland J (1996) Demographic Data Collection in Less Developed Countries 1946-1996. Population Studies 50: 433-450.

14. Greenfield C C (1976) A Revised Procedure for Dual Record Systems in Estimating Vital Events. Jour. Royal Statist. Society A 139(3): 389-401.

15. Sabagh G, C Scott (1967) A Comparison of Different Survey techniques for Obtaining Vital data in a Developing Country. Demography 4: 759-772.