

A Simulation Study on the Effect of Matching Error on Triple-System Estimation

Richard A Griffin

U.S. Census Bureau, Washington, DC

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

1. Background

Residual heterogeneity (variation in capture probabilities) is known to produce bias in the dual-system estimates which have been used to estimate census coverage in U.S. Censuses since 1980. Triple-system estimation (TSE) using an administrative records list as a third source along with the census and post-enumeration survey (PES) has the potential to produce estimates with less bias.

Griffin (2013) presented and evaluated potential statistical methods for estimation of net census undercount using three systems for obtaining population information: (1) a decennial census; (2) an independent enumeration of the population; and (3) administrative records. This paper was prepared for the 2012 American Statistical Association Hard to Reach Conference. Results showed that three sets of capture attempts can produce more accurate estimates than two capture attempts. However, the paper noted that increased matching error, which was ignored, seems likely going from the two capture attempts necessary for dual-system estimation (DSE) to the three capture attempts necessary for TSE. For two attempts at capture, there are only four cells in a 2×2 cross classification of capture status and given the marginal counts of the total for each of the attempts, matching is only necessary to obtain the cell representing captured in both attempts. For three attempts, there are eight cells in a $2 \times 2 \times 2$ cross classification of capture status. Obtaining all these counts from a complex matching operation may be error prone. This paper provides a simulation to investigate the effect of matching error on DSE as well as on some of the potential triple-system estimates possible if an administrative records system is added to a census followed by a PES.

The incomplete 2^3 table of counts for TSE can be divided into one complete 2×2 sub-table and one incomplete 2×2 sub-table. The additional source from administrative records provides data with which to evaluate the previously un-testable assumption of independence between the census and the PES. Direct evidence is available in the triple-system tables for odds ratios in 2×2 sub-tables formed by restricting consideration to cases observed in the administrative records source. In this case, complete information is available for all four cells defined by capture or not in the census and PES. This additional information is used to formulate the triple-system estimates using an assortment of model assumptions.

It is assumed that all N individuals in the population are exposed to possible inclusion in all three sources. In practice, sampling is necessary for the PES and possibly the administrative list (due to the necessity of follow-up for unresolved match status). In addition, erroneous inclusions including within-list duplicates have been removed from all lists. The model assumes autonomous independence, which means that the census list, the PES list and the administrative list are created as a result of N mutually independent trials from one person to the next (all persons are captured independently of all other persons).

Section 2 provides a matching error model for DSE and section 3 expands this to a matching error model applicable for TSE. Section 4 describes three alternative estimators for TSE that are compared with each other and with DSE in the simulations. Sections 5 and 6 give the details for creating the simulated populations and replication of the simulations. This is followed by sections 7, 8, and 9 providing results, analysis, summary, and conclusions.

1. A Matching Error Model for Dual-System Estimation

Biemer (1988) presents a model for the matching error associated with DSE . Here we present portions of Biemer’s model as background for the TSE matching error model. The motivation is that although a TSE may reduce model errors inherent in DSE, the additional matching error associated with TSE may counter balance this gain and result in an increase in total error.

For DSE, there are two list sources, a census and an independent PES. Table 1 shows the observed counts after matching persons from the PES to the census to obtain X_{11} .

Table 1: Observed Counts after Matching for Dual-System Estimation

	In PES 1	Out of PES 0	
In Census 1	X_{11}	$X_{1+} - X_{11}$	$X_{1+} = N_C$
Out of Census 0	$X_{+1} - X_{11}$		
	$X_{+1} = N_P$		N

Matching error has no effect on the census count, N_C , or the PES count, N_P .

The true population count, N , is unknown. For all persons in the PES, define a 1 or 0 indicator, y , as 1 for matched and 0 for not matched to the census. In addition, define a 1 or 0 indicator, z , as 1 for truly in the census and 0 for truly missed in the census.

Assume that matching is independent from person to person and define matching error for person j as follows:

$$P\{y_j = 1 \mid z_j = 0\} = \phi \text{ (probability of false positive)}$$

$$P\{y_j = 0 \mid z_j = 1\} = \alpha \text{ (probability of false negative)}$$

These error probabilities are assumed to be homogeneous across persons enumerated in the PES.

Define $p_{11} = \frac{\sum_{j=1}^{N_p} z_j}{N_p}$ and $X_{11} = \sum_{j=1}^{N_p} y_j$. p_{11} is the true proportion of persons in the PES who are also in the census.

X_{11} is the observed count after matching of persons in the PES who are matched to the census.

Then the expected value of X_{11} is given as follows:

$$E(X_{11}) = \sum_{j=1}^{N_p} E(y_j) = N_p p_{11} (1 - \alpha) + N_p (1 - p_{11}) \phi.$$

As for variance, $V(y_j | z_j = 0) = \phi(1 - \phi)$ and $V(y_j | z_j = 1) = (1 - \alpha)\alpha$ so that

$$V(X_{11}) = \sum_{j=1}^{N_p} V(y_j) = N_p p_{11} \alpha (1 - \alpha) + N_p (1 - p_{11}) \phi (1 - \phi).$$

$$\text{Define } \hat{p}_{11} = \frac{\sum_{j=1}^{N_p} y_j}{N_p}.$$

\hat{p}_{11} is the estimate of p_{11} using the results of the matching operation.

Then the dual-system estimate, which assumes independence between the PES and Census is given by $I = \frac{N_C}{\hat{p}_{11}}$.

Next

$$E(\hat{p}_{11}) = p_{11} (1 - \alpha) + (1 - p_{11}) \phi = p_{11}^* \text{ (for notational purposes) and}$$

$$V(\hat{p}_{11}) = \frac{1}{N_p} [p_{11} (1 - \alpha) \alpha + (1 - p_{11}) \phi (1 - \phi)].$$

Using Taylor series expansions (first and second partial derivative terms for expected value and only first partial derivative term for variance),

$$E(I) \cong \frac{N_C}{p_{11}^*} + \frac{N_C}{(p_{11}^*)^3} V(\hat{p}_{11}) \text{ and}$$

$$V(I) \cong \frac{N_C^2}{(p_{11}^*)^4} V(\hat{p}_{11}).$$

2. A Matching Error Model for Triple-System Estimation

The three systems for obtaining population information for TSE used for this paper are (1) a decennial census; (2) an independent enumeration of the population; and (3) administrative records. There are many possible estimators associated with TSE. Griffin (2013) documents and uses ten estimators in a simulation. Seven of these estimators are motivated by hierarchical log linear models based on Fienberg (1972). Two of the estimators are based on suggestions from Zaslavsky and Wolfgang (1990 and 1993). Other estimators are provided in Darrock et. al (1993).

The observed counts after matching for a triple-system framework are shown in Table 2.

Table 2: Observed Counts after matching for Triple-System Estimation

	In Administrative List		Out of Administrative List	
	In PES 1	Out of PES 0	In PES 1	Out of PES 0
In Census 1	X_{111}	X_{101}	X_{110}	X_{100}
Out of Census 0	X_{011}	X_{001}	X_{010}	

The matching associated with TSE is much more complicated than that for DSE. Several possible strategies can be envisioned. For a TSE matching error model, assume the following matching procedure is used to obtain the counts in Table 2. Define three sets containing all persons enumerated on a list. P is all persons enumerated in the PES, C is all persons enumerated in the Census, and A is all persons enumerated on the administrative list.

Step 1: First match P to C and then P to A to determine X_{111} , X_{110} , X_{011} , and X_{010}

Step 2: Next match A to C to determine X_{101} and X_{001}

Step 3: Finally, match C to A and P to determine X_{100}

Each of these matching steps is assumed to be independent. Thus no changes to a count obtained in step 1 or step 2 occur as a result of a later step.

The first estimator considered for this paper is the estimator associated with the **no-second-order-interaction** log linear model. This is the least restrictive log linear model for which data is available for estimation. The incomplete 2^3 table of counts in Table 2 is divided into one complete 2×2 sub-table and one incomplete sub-table. Assume the cross-product ratio is the same in both sub-tables. Then the estimate of the missing cell in the incomplete 2×2 table can be estimated using the known cross-product ratio from the complete 2×2 table. The assumption is that the dependence in the 2×2 table for $C \times P$ using only those individuals in A , is the same as the dependence in the 2×2 table for $C \times P$ using only those individuals not in A . This model is in some sense analogous to the

assumption of independence for the 2×2 table used for DSE but is one layer deeper. All pairs of sources can exhibit dependence, but the amount of dependence in each pair is assumed to be unaffected by conditioning on the third source.

The estimator for this model is

$$NSOI = X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} + \frac{(X_{111})(X_{001})(X_{100})(X_{010})}{(X_{011})(X_{101})(X_{110})}.$$

Matching error has no effect on the census count, N_C , the PES count, N_P , or the administrative list count, N_A .

The true population count, N , is unknown as is the true population count in each cell denoted by N_{ijk} . For example the true count in the 101 cell is N_{101} while the observed count after matching is X_{101} .

For all persons in P or A, define a 1 or 0 indicator, y , as 1 for matched and 0 for not matched to the census.

For all persons in P or C, define a 1 or 0 indicator, w , as 1 for matched and 0 for not matched to the administrative list.

For all persons in C or A, define a 1 or 0 indicator, z , as 1 for matched and 0 for not matched to the PES.

In addition, define a 1 or 0 indicator, c , as 1 for truly in the census and 0 for truly missed in the census.

Define a 1 or 0 indicator, a , as 1 for truly on the administrative list and 0 for truly missed on the administrative list.

Define a 1 or 0 indicator, p , as 1 for truly in the PES and 0 for truly missed in the PES.

With this notation, the observed counts after matching are as follows:

$$\begin{aligned} X_{111} &= \sum_{j=1}^{N_P} y_j w_j; & X_{110} &= \sum_{j=1}^{N_P} y_j (1 - w_j); & X_{011} &= \sum_{j=1}^{N_P} (1 - y_j) w_j; & X_{010} &= \sum_{j=1}^{N_P} (1 - y_j) (1 - w_j); \\ X_{011} &= \sum_{j=1}^{N_A} y_j (1 - z_j); & X_{001} &= \sum_{j=1}^{N_A} (1 - y_j) (1 - z_j); & X_{001} &= \sum_{j=1}^{N_C} (1 - z_j) (1 - w_j) \end{aligned}$$

Expected Value and Variance

Expressions for the expected value and variance of these observed counts as well as the expected value and variance of NSOI using Taylor Linearization are provided in Appendix 1.

3. Alternative Estimators

The no second order interaction estimator, NSOI, requires the observed counts after matching for each of the seven observed cells from Table 2. Thus, NSOI may have a large bias if there is substantial matching error. As, mentioned earlier, several alternative estimators using three lists are presented in Griffin (2013). Seven of these estimators are motivated by hierarchical log linear models as described in Fienberg (1972). The no-second-order-interaction model {CP, PA, CA} is the least restrictive log linear model for which data is available for estimation. Three of the seven estimators are based on conditionally independent log linear models and three others are based on jointly independent log linear models. In each case, the three estimators are of the same format based on the dependence between sources C, P, and A taken two sources at a time. Here we consider one conditionally independent model and one jointly independent model.

Jointly independent model {A, CP} assumes that there is a relationship between C and P, but neither C nor P has a relationship with A. This is ordinary two-way independence between A and a categorical variable composed of all four combinations (captured/not captured) of C and P. The estimate for the 000 cell is the usual dual-system estimate for the unobserved cell in the 2×2 table for which one list is A and the other list is formed by combining C and P (un-duplication requiring matching is necessary). The population estimator for this model is as follows:

$$JI = X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} + \frac{(X_{001})(X_{110} + X_{100} + X_{010})}{(X_{111} + X_{101} + X_{011})}.$$

Conditionally independent model {CP, CA} assumes that each level (captured/not captured) of C, P and A are independent. The estimate for the 000 cell is the usual dual-system estimate for the unobserved cell in the 2×2 table conditional on C = 0, using the A list and the P list after removing all individuals captured on the C list (un-duplication requiring matching is necessary). The population estimator for this model is as follows:

$$CI = X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} + \frac{(X_{001})(X_{010})}{(X_{011})}.$$

It seems possible that either of these alternative estimators could be less affected by matching error than TSE. The estimate for the 000 cell for CI only uses three of the seven observed counts after matching. The estimate for the 000 cell for JI uses all observed counts but sums of observed counts may have less net error than the total error over each individual cell.

Expressions for the expected value and variance of JI using Taylor Linearization are provided in Appendix 2.

Expressions for the expected value and variance of CI using Taylor Linearization are provided in Appendix 3.

4. Creating the Simulated Populations

Populations of N = 1000 persons will be simulated, allowing for heterogeneous capture probabilities and homogeneous conditional odds ratios. One conditional odds ratio, is the odds ratio for the 2×2 table of CxP conditional on capture on A and the other is the odds ratio for the 2×2 table of CxP conditional on not captured (missed) on A.

4.1 Creating a Specified Conditional Odds Ratio

Omitting any subscript for an individual member of the population, the 2×2 table of conditional capture probabilities for census capture and PES capture given capture on the administrative list is given in Table 3.

Table 3: Capture Probabilities for Census and PES given Capture on Administrative list

	In PES 1	Out of PES 0	
In Census 1	P_{11}	P_{10}	P_{1+}
Out of Census 0	P_{01}	P_{00}	
	P_{+1}		

In order to create a simulated population with a given set of conditional odds ratios, the odds ratio formula for a 2×2 sub-table is written as a function of an unknown proportion in the 11 cell (P_{11}) and the known marginal proportions (P_{1+} and P_{+1}).

$$\text{Thus given } P_{1+}, P_{+1}, \text{ and odds ratio } \theta = \frac{P_{11}P_{00}}{P_{10}P_{01}} = \frac{P_{11}(1 - P_{1+} - P_{+1} + P_{11})}{(P_{1+} - P_{11})(P_{+1} - P_{11})},$$

$$\text{The equation can be re-written as } (1 - \theta)P_{11}^2 + [1 - P_{1+} - P_{+1} + \theta(P_{1+} + P_{+1})]P_{11} - \theta P_{1+}P_{+1} = 0. \quad (1)$$

This equation can be solved for P_{11} using the quadratic formula producing two roots one of which is between 0 and 1 and is the one we want.

This value of P_{11} and given P_{1+} and P_{+1} provides the desired odds ratio θ .

The process described starting with Table 3 is repeated for capture probabilities for the census and the PES given not captured (missed) on the administrative list allowing in some simulations for a different conditional odds ratio θ .

5.2 Generating a 1000 Person Population Allowing for Heterogeneity in Capture Probabilities

We want to generate several populations of size $N = 1000$ persons to have particular capture properties. This is accomplished by specifying two conditional odds ratios.

Let θ_1 be the odds ratio for census and PES given capture on the administrative list, and θ_2 be the odds ratio for census and PES given **not** captured on the administrative list.

Given θ_1 and θ_2 (assumed constant over persons) and ten beta parameters in the following conditional capture probabilities

$$P_k \langle A \rangle = \frac{\exp(\beta_{10} + \beta_{11}X_k)}{1 + \exp(\beta_{10} + \beta_{11}X_k)}, P_k \langle C|A \rangle = \frac{\exp(\beta_{20} + \beta_{21}X_k)}{1 + \exp(\beta_{20} + \beta_{21}X_k)}, P_k \langle P|A \rangle = \frac{\exp(\beta_{30} + \beta_{31}X_k)}{1 + \exp(\beta_{30} + \beta_{31}X_k)},$$

$$P_k \langle C|notA \rangle = \frac{\exp(\beta_{40} + \beta_{41}X_k)}{1 + \exp(\beta_{40} + \beta_{41}X_k)}, P_k \langle P|notA \rangle = \frac{\exp(\beta_{50} + \beta_{51}X_k)}{1 + \exp(\beta_{50} + \beta_{51}X_k)},$$

for $k = 1$ to 1000, independently generate $X_k \sim N(0,1)$ and calculate

$$P_k \langle A \rangle, P_k \langle C|A \rangle, P_k \langle P|A \rangle, P_k \langle C|notA \rangle, P_k \langle P|notA \rangle.$$

Note that although the conditional odds ratios are assumed constant over persons, the capture probabilities are heterogeneous since variation in the independent variables is created.

Using θ_1 and $P_k \langle C|A \rangle, P_k \langle P|A \rangle$, we use the methodology from section 5.1 and equation (1) to solve for the probability of capture in both the census and PES given capture on the administrative list. Then complete the 2×2

table of capture probabilities given capture on the administrative list. Multiplying each of these conditional probabilities by $P_k \langle A \rangle$ provides $p_{k,111}, p_{k,101}, p_{k,011}, p_{k,001}$

Then, using θ_2 and $P_k \langle C | \text{not} A \rangle, P_k \langle P | \text{not} A \rangle$, use the methodology from section 5.1 and equation (1) to solve for the probability of capture in both the census and PES given **not** captured on the administrative list. Then complete the 2x2 table of capture probabilities given **not** captured on the administrative list. Multiplying each of these conditional probabilities by $(1 - P_k \langle A \rangle)$ provides $p_{k,110}, p_{k,100}, p_{k,010}, p_{k,000}$.

Next, generate a number u from 0 to 1 from the distribution Uniform (0,1) and use the cumulative distribution of the eight cell probabilities to determine which of the eight cells of Table 2 person k falls

After completing the above for each of the 1000 population persons, tabulate the seven observed counts from Table 2 and using these compute $R^t = \frac{E(t)}{1000}$ for $t = \text{NSOI, JI, and CI}$. This is the ratio of the expected value (using the expressions given in the Appendix) of the estimated population count to the true population count and provides a measure of the accuracy of the estimate.

5. Replication

This paper presents results for 1000 independent replications of the population generation as specified in section 5.2 for a given θ_1 and θ_2 (assumed constant over persons) and one set of beta parameters.

For each of the three model estimates $t = \text{NSOI, JI and CI}$, use these 1000 replicates to compute the empirical mean ratio R^t denoted as \bar{R}^t , and its variance, $\text{Var}(\bar{R}^t)$.

Note that none of the precise assumptions, particularly homogeneity in capture probability, needed for validity of any of these ten estimators is satisfied by any of these simulated populations. Darroch et al. (1993) provide some arguments that the no three-way interaction model may be a fair approximation except for heterogeneity. The kind of person-to-person heterogeneity introduced by these simulations might be expected to be a reasonable representation of the reality of list formation. This heterogeneity produces bias in these estimates even if the model assumptions about the relationship between the capture attempts hold.

6. Results

Define the ‘‘Average Capture Probability’’ (ACP) as the average of the five probabilities defined in section 5.2 for $X_k = 0$ (the mean of the random variable X). It is used as a measure of ‘‘Hard to Reach’’ since lower values indicate lower capture probabilities (i.e., harder to reach).

$$ACP = \frac{\sum_{i=1}^5 \frac{e^{\beta_{i0}}}{1 + e^{\beta_{i0}}}}{5}.$$

Table 4 shows results for each of the four estimator alternatives for four sets of odds ratios θ_1 and θ_2 (1.5 and 1.2; 0.75 and 0.85; 0.75 and 0.75; 1.5 and 1.5.) using a false positive rate of 0.062 and a false negative rate of 0.023. These error rates are from Biemer (1988) and are for a computer matching operation from a 1986 pretest of PES matching procedures in Los Angeles, CA. When $\theta_1 = \theta_2$, the odds ratio for census capture or not by PES capture status is independent of capture status on the administrative list (no second order interaction). When, $\theta_1 \neq \theta_2$ the

odds ratio for census capture or not by PES capture status is dependent on capture status on the administrative list. For each estimator the mean ratio of the expected value of the estimated population count to the true population count over the thousand replicates, \bar{R}^t , and its standard error, $Var(\bar{R}^t)$ are shown.

The mean ratio results vary by odds ratio alternatives. For $\theta_1 = 1.5$ and $\theta_2 = 1.2$, estimator I (using DSE) was best (closest to 1) with an average R of 1.002 (se = 0.003) and the best triple-system estimator was CI with an average R of 0.963 (se = 0.002). For $\theta_1 = 0.75$ and $\theta_2 = 0.85$, estimator I (using DSE) had an average R of 1.283 (se = 0.005) and the best triple-system estimator was NSOI with an average R of 1.015 (se = 0.005). For $\theta_1 = 0.75$ and $\theta_2 = 0.75$, estimator I (using DSE) had an average R of 1.332 (se = 0.005) and the best triple-system estimator was CI with an average R of 0.974 (se = 0.002). For $\theta_1 = 1.5$ and $\theta_2 = 1.5$, estimator I (using DSE) had an average R of 0.934 (se = 0.003) and the best triple-system estimator was NSOI with an average R of 1.052 (se = 0.005).

Table 5 shows the average standard error of the alternative estimators for the population of true size 1000. The average of these averages over the odds ratio alternatives are about 97 for NSOI, 24 for JI, 21 for CI, and 34 for I (using DSE).

8. Analysis

Doing this simulation with different β parameters and odds ratio alternatives (results from these alternatives not presented in this paper) has shown that results vary. In particular the excellent performance of DSE for $\theta_1 = 1.5$ and $\theta_2 = 1.2$ (average R of 1.002), was observed for only that set of parameters and is not necessarily an indication that DSE is in general less biased when matching error is incorporated in bias measures.

The odds ratio for census and PES given capture on the administrative list may well be greater than the odds ratio for census and PES given **not** captured on the administrative list ($\theta_1 > \theta_2$). It may also be likely that both of these odds ratios are greater than 1. Keeping in mind that results will vary for different β parameters and odds ratio alternatives, Table 6 shows some results for the same β parameters and for $\theta_1 = 1.75$ and $\theta_2 = 1.5$ allowing the false positive and false negative error rates to vary. The first set of error rates are 0.062 and 0.023 as in Table 4. Error rates about twice as large (0.12 and 0.05) and about half as large (0.03 and 0.01) are also shown as is the cases of no matching error at all. For error rates of 0.062 and 0.023, DSE had an average R of 0.958 and the best triple-system estimator was CI with an average R of 0.967. For error rates about twice as large (0.12 and 0.05), DSE had an average R of 0.958 and the best triple-system estimator was CI with an average R of 0.948. For error rates about half as large (0.03 and 0.01), DSE had an average R of 0.967 and the best triple-system estimator was CI with an average R also of 0.967. For no matching error, DSE had an average R of 0.952 and the best triple-system estimator was CI with an average R of 0.965.

9. Summary and Conclusion

Three sets of capture attempts can produce more accurate estimates than two capture attempts. However, there is likely to be increased matching error going from two attempts to three attempts. For two attempts at capture, there are only four cells in a 2×2 table. Given the marginal counts of the total count for each of the attempts, matching is only necessary to obtain the 11 cell (captured in both attempts). For three attempts, there are eight cells. For NSOI, no-second-order-interaction, counts are required for all observable seven cells in order to estimate the 000 cell. NSOI makes a less restrictive assumption (no second order interaction) than CI (conditionally independent) and JI (jointly independent). In theory, with no matching error, NSOI should be the better than JI, CI, or I (using DSE) if there are no other errors in obtaining the counts. Second order interaction and heterogeneity in capture probabilities are likely in the real world for most populations. For example, both the 111 cell and the 110 cell are required for TSE so that both the count of captured in the first two attempts and in the third attempt AND captured in the first

two attempts but missed in the third are necessary. Obtaining all these counts from a complex matching operation may be error prone.

Since doing this simulation with different β parameters and alternative odds ratios (not presented in this paper) has shown that results vary, the conclusions given in the paper serve as an illustration of potential results and cannot be used to make definitive generalizations. The results shown indicate that that matching error can lessen the theoretical advantages of triple-system modeling. Due to matching error, the DSE may be more accurate than TSE using any of the triple-system estimators. For the triple-system estimators, CI or JI may be more accurate than NSOI due to using fewer of the seven observed cells in the triple-system setup. Increasing the matching error from the levels used by Biemer (1988) has a great effect on the accuracy of all the estimators. The best estimators, CI and DSE, for the Biemer level matching error had about a 4 percent undercount and doubling the matching error resulted in about the same accuracy. Decreasing the matching error level produced little change in accuracy. The relative accuracy among the estimators did not change much with varying matching error levels.

References

- Alho, J. (1990), "Logistic Regression in Capture-Recapture Models," *Biometrics*, 46, 623-635.
- Bell, W. (1993), "Using Information from Demographic Analysis in Post-Enumeration survey Estimation," *Journal of the American Statistical Association*, 88, 1106-1118.
- Biemer, P. (1988), "Using Information from Demographic Analysis in Post-Enumeration survey Estimation," *Survey Methodology*, 14, 117-134
- Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993), "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability," *Journal of the American Statistical Association*, 88, 1137-1148.
- Fienberg, S. (1972), "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables," *Biometrika*, Vol. 59, No.3, 591-603.
- Griffin, R. (2013), "Potential Uses of Administrative Records for Tripe System Modeling for Estimation of Census Coverage Error in 2020", *Proceedings of the American Statistical Association 2012 Hard to Reach Conference*.
- Zaslavsky, A. and Wolfgang, G. (1990), "Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative List Data," *Proceedings of the Survey Research Methods Section of the American Statistical Association*.
- Zaslavsky, A. and Wolfgang, G. (1993), "Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative List Data," *Journal of Business & Economic Statistics*, 11, 279-288.

Table 4: Accuracy of Alternative Estimators by Varying Odds Ratios

False Positive Rate = 0.062								
False Negative Rate = 0.023								
Odds Ratios	$\Theta_1 = 1.5$ $\Theta_2 = 1.2$		$\Theta_1 = 0.75$ $\Theta_2 = 0.85$		$\Theta_1 = 0.75$ $\Theta_2 = 0.75$		$\Theta_1 = 1.5$ $\Theta_2 = 1.5$	
Estimator	\bar{R}'	$SE(\bar{R}')$	\bar{R}'	$SE(\bar{R}')$	\bar{R}'	$SE(\bar{R}')$	\bar{R}'	$SE(\bar{R}')$
NSOI	1.124	0.006	1.015	0.005	1.047	0.006	1.052	0.005
JI	1.088	0.002	1.080	0.002	1.089	0.002	1.079	0.002
CI	0.963	0.002	0.955	0.002	0.974	0.002	0.946	0.002
DSE	1.002	0.003	1.283	0.005	1.332	0.006	0.934	0.003

Table 5: Average Standard Error of Estimates by Varying Odds Ratios

False Positive Rate = 0.062				
False Negative Rate = 0.023				
Odds Ratio	$\Theta_1 = 1.5$ $\Theta_2 = 1.2$	$\Theta_1 = 0.75$ $\Theta_2 = 0.85$	$\Theta_1 = 0.75$ $\Theta_2 = 0.75$	$\Theta_1 = 1.5$ $\Theta_2 = 1.5$
Estimator	Standard Error	Standard Error	Standard Error	Standard Error
NSOI	106	93	100	90
JI	25	24	24	24
CI	22	21	21	21
DSE	32	36	36	31

Table 6: Accuracy of Alternative Estimators by Varying Error Rates

Fixed Odds Ratios	$\Theta_1 = 1.75$ $\Theta_2 = 1.25$			
False Positive Error Rate	0.062	0.12	0.03	0
False Negative Error Rate	0.023	0.05	0.01	0
	\bar{R}'	\bar{R}'	\bar{R}'	\bar{R}'
NSOI	1.183	1.121	1.220	1.125
JI	1.093	1.066	1.106	1.098
CI	0.967	0.948	0.967	0.965
DSE	0.958	0.958	0.967	0.952

Appendix 1: Expected Value and Variance of No-Second-Order-Interaction Model

For the observed seven counts after matching, as an example, all the details are first provided for X_{110} .

For any person j ,

$y_j(1-w_j) = 1$ if $y_j=1$ and $w_j=0$, otherwise $y_j(1-w_j) = 0$.

Thus $y_j(1-w_j)$ is a Bernoulli random variable with parameter $p_j = \Pr(y_j = 1)\Pr(w_j = 0)$.

$$\text{Thus } E(X_{110}) = \sum_{j=1}^{N_p} p_j \text{ and } V(X_{110}) = \sum_{j=1}^{N_p} p_j(1-p_j).$$

Persons enumerated in the PES can truly be in cells 111, 110, 011, or 010. p_j is different for each of these true cells.

Using the cell subscripts, for this example, person j is truly in cell 010, $p_j = p_{010}$.

Thus,

$$p_{111} = \Pr(y_j = 1 | c_j = 1)\Pr(w_j = 0 | a_j = 1) = (1 - \alpha)\alpha$$

$$p_{110} = \Pr(y_j = 1 | c_j = 1)\Pr(w_j = 0 | a_j = 0) = (1 - \alpha)(1 - \phi)$$

$$p_{011} = \Pr(y_j = 1 | c_j = 0)\Pr(w_j = 0 | a_j = 1) = \phi\alpha$$

$$p_{010} = \Pr(y_j = 1 | c_j = 0)\Pr(w_j = 0 | a_j = 0) = \phi(1 - \phi)$$

So that

$$E(X_{110}) = N_{111}p_{111} + N_{110}p_{110} + N_{011}p_{011} + N_{010}p_{010} \text{ and}$$

$$V(X_{110}) = N_{111}p_{111}(1-p_{111}) + N_{110}p_{110}(1-p_{110}) + N_{011}p_{011}(1-p_{011}) + N_{010}p_{010}(1-p_{010})$$

Using this procedure,

$$E(X_{111}) = N_{111}p_{111} + N_{110}p_{110} + N_{011}p_{011} + N_{010}p_{010}$$

$$V(X_{111}) = N_{111}p_{111}(1-p_{111}) + N_{110}p_{110}(1-p_{110}) + N_{011}p_{011}(1-p_{011}) + N_{010}p_{010}(1-p_{010}), \text{ where}$$

$$p_{111} = (1 - \alpha)^2; p_{110} = (1 - \alpha)\phi; p_{011} = \phi(1 - \alpha); p_{010} = \phi^2$$

$$E(X_{011}) = N_{111}p_{111} + N_{110}p_{110} + N_{011}p_{011} + N_{010}p_{010}$$

$$V(X_{011}) = N_{111}p_{111}(1-p_{111}) + N_{110}p_{110}(1-p_{110}) + N_{011}p_{011}(1-p_{011}) + N_{010}p_{010}(1-p_{010}), \text{ where}$$

$$p_{111} = \alpha(1 - \alpha); p_{110} = \alpha\phi; p_{011} = (1 - \phi)(1 - \alpha); p_{010} = (1 - \phi)\phi$$

$$E(X_{010}) = N_{111}p_{111} + N_{110}p_{110} + N_{011}p_{011} + N_{010}p_{010}$$

$$V(X_{010}) = N_{111}p_{111}(1-p_{111}) + N_{110}p_{110}(1-p_{110}) + N_{011}p_{011}(1-p_{011}) + N_{010}p_{010}(1-p_{010}), \text{ where } p_{111} = \alpha^2; p_{110} = \alpha(1-\phi); p_{011} = (1-\phi)\alpha; p_{010} = (1-\phi)^2$$

$$E(X_{101}) = N_{111}p_{111} + N_{101}p_{101} + N_{011}p_{011} + N_{001}p_{001}$$

$$V(X_{101}) = N_{111}p_{111}(1-p_{111}) + N_{101}p_{101}(1-p_{101}) + N_{011}p_{011}(1-p_{011}) + N_{001}p_{001}(1-p_{001}), \text{ where } p_{111} = (1-\alpha)\alpha; p_{101} = (1-\alpha)(1-\phi); p_{011} = \phi\alpha; p_{001} = \phi(1-\phi)$$

$$E(X_{001}) = N_{111}p_{111} + N_{101}p_{101} + N_{011}p_{011} + N_{001}p_{001}$$

$$V(X_{001}) = N_{111}p_{111}(1-p_{111}) + N_{101}p_{101}(1-p_{101}) + N_{011}p_{011}(1-p_{011}) + N_{001}p_{001}(1-p_{001}), \text{ where } p_{111} = \alpha^2; p_{101} = \alpha(1-\phi); p_{011} = (1-\phi)\alpha; p_{001} = (1-\phi)^2$$

$$E(X_{100}) = N_{111}p_{111} + N_{110}p_{110} + N_{101}p_{101} + N_{100}p_{100}$$

$$V(X_{100}) = N_{111}p_{111}(1-p_{111}) + N_{110}p_{110}(1-p_{110}) + N_{101}p_{101}(1-p_{101}) + N_{100}p_{100}(1-p_{100}), \text{ where } p_{111} = \alpha^2; p_{110} = \alpha(1-\phi); p_{101} = (1-\phi)\alpha; p_{100} = (1-\phi)^2$$

Covariance terms necessary for the TAYLOR Linearization variance approximation

Due to the assumptions of independence between matching steps, there are seven non-zero covariance terms among the observed counts after matching.

From Step 1:

$$\text{cov}(X_{111}, X_{110}); \quad \text{cov}(X_{111}, X_{011}); \quad \text{cov}(X_{111}, X_{010});$$

$$\text{cov}(X_{110}, X_{011}); \quad \text{cov}(X_{110}, X_{010}); \quad \text{cov}(X_{011}, X_{010})$$

From Step 2: $\text{cov}(X_{101}, X_{001})$

From Step 3: There are no covariance terms since only one count is observed from step 3 matching.

Details of the derivation are provided for $\text{cov}(X_{111}, X_{110})$.

$$\text{cov}(X_{111}, X_{110}) = \text{cov}\left(\sum_{j=1}^{N_p} y_j w_j, \sum_{j=1}^{N_p} y_j (1-w_j)\right)$$

Let $a_j = y_j w_j$ and $b_j = y_j (1-w_j)$. $a_j = 1$ then $b_j = 0$.

$$\begin{aligned} \text{cov}\left(\sum_{j=1}^{N_P} a_j, \sum_{j=1}^{N_P} b_j\right) &= \sum_{j=1}^{N_P} \text{cov}(a_j, b_j) + \sum_{j \neq i}^{N_P} \text{cov}(a_j, b_i) = \sum_{j=1}^{N_P} [E(a_j b_j) - E(a_j)E(b_j)] + 0 \\ &= -\Pr(a_j = 1)\Pr(b_j = 1) \end{aligned}$$

Persons enumerated in the PES can truly be in cells 111, 110, 011, or 010.

For cell 111.

$$p_{111} = \Pr(y_j = 1 | c_j = 1)\Pr(w_j = 1 | a_j = 1) = (1 - \alpha)^2$$

$$p_{110} = \Pr(y_j = 1 | c_j = 1)\Pr(w_j = 1 | a_j = 0) = (1 - \alpha)\phi$$

$$p_{011} = \Pr(y_j = 1 | c_j = 0)\Pr(w_j = 1 | a_j = 1) = \phi(1 - \alpha)$$

$$p_{010} = \Pr(y_j = 1 | c_j = 0)\Pr(w_j = 1 | a_j = 0) = \phi^2$$

For cell 110

$$p_{111} = \Pr(y_j = 1 | c_j = 1)\Pr(w_j = 0 | a_j = 1) = (1 - \alpha)\alpha$$

$$p_{110} = \Pr(y_j = 1 | c_j = 1)\Pr(w_j = 0 | a_j = 0) = (1 - \alpha)(1 - \phi)$$

$$p_{011} = \Pr(y_j = 1 | c_j = 0)\Pr(w_j = 0 | a_j = 1) = \phi\alpha$$

$$p_{010} = \Pr(y_j = 1 | c_j = 0)\Pr(w_j = 0 | a_j = 0) = \phi(1 - \phi)$$

$$\text{Thus } \text{cov}(X_{111}, X_{110}) = N_{111}(1 - \alpha)^3 \alpha + N_{110}(1 - \alpha)^2 \phi(1 - \phi) + N_{011}\phi^2(1 - \alpha)\theta + N_{010}\phi^3(1 - \phi)$$

The other non-zero covariance terms are as follows:

$$\text{cov}(X_{111}, X_{011}) = N_{111}(1 - \alpha)^3 \alpha + N_{110}(1 - \phi)\phi^2 \alpha + N_{011}\phi(1 - \phi)(1 - \alpha) + N_{010}\phi^3(1 - \phi)$$

$$\text{cov}(X_{111}, X_{010}) = N_{111}(1 - \alpha)^2 \alpha^2 + N_{110}(1 - \theta)^2 \phi\alpha + N_{011}\phi\alpha(1 - \alpha)(1 - \phi) + N_{010}\phi^2(1 - \phi)^2$$

$$\text{cov}(X_{110}, X_{011}) = N_{111}(1 - \alpha)^2 \alpha^2 + N_{110}(1 - \alpha)\alpha\phi(1 - \phi) + N_{011}\phi(1 - \phi)(1 - \alpha)\alpha + N_{010}\phi^2(1 - \phi)^2$$

$$\text{cov}(X_{110}, X_{010}) = N_{111}(1 - \alpha)\alpha^3 + N_{110}(1 - \alpha)\alpha(1 - \phi)^2 + N_{011}\phi(1 - \phi)\alpha^2 + N_{010}\phi(1 - \phi)^3$$

$$\text{cov}(X_{011}, X_{010}) = N_{111}(1 - \alpha)\alpha^3 + N_{110}\alpha^2\phi(1 - \phi) + N_{011}(1 - \phi)^2(1 - \alpha)\alpha + N_{010}\phi(1 - \phi)^3$$

$$\text{cov}(X_{101}, X_{001}) = N_{111}(1 - \alpha)\alpha^3 + N_{110}(1 - \alpha)\alpha(1 - \phi)^2 + N_{011}\phi(1 - \phi)\alpha^2 + N_{010}\phi(1 - \phi)^3$$

Taylor Linearization

For any of the seven observed counts after matching, let $E(X_{ijk}) = E_{ijk}$.

The estimator of population, NSOI, is approximated using Taylor Linearization about the vector of expected values of the seven observed counts resulting in,

$$\begin{aligned}
NSOI &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} + \frac{(E_{111})(E_{001})(E_{100})(E_{010})}{(E_{011})(E_{101})(E_{110})} \\
&+ \frac{E_{001}E_{100}E_{010}}{E_{011}E_{101}E_{110}}(X_{111} - E_{111}) + \frac{E_{111}E_{100}E_{010}}{E_{011}E_{101}E_{110}}(X_{001} - E_{001}) + \frac{E_{111}E_{001}E_{010}}{E_{011}E_{101}E_{110}}(X_{100} - E_{100}) + \frac{E_{111}E_{001}E_{100}}{E_{011}E_{101}E_{110}}(X_{010} - E_{010}) \\
&- \frac{E_{111}E_{001}E_{100}E_{010}}{E_{011}^2E_{101}E_{110}}(X_{011} - E_{011}) - \frac{E_{111}E_{001}E_{100}E_{010}}{E_{101}^2E_{011}E_{110}}(X_{101} - E_{101}) - \frac{E_{111}E_{001}E_{100}E_{010}}{E_{110}^2E_{101}E_{011}}(X_{110} - E_{110}) \\
&+ \frac{1}{2} \left[\frac{2E_{111}E_{001}E_{100}E_{010}}{E_{011}^3E_{101}E_{110}}(X_{011} - E_{011})^2 + \frac{2E_{111}E_{001}E_{100}E_{010}}{E_{101}^3E_{011}E_{110}}(X_{101} - E_{101})^2 + \frac{2E_{111}E_{001}E_{100}E_{010}}{E_{110}^3E_{101}E_{011}}(X_{110} - E_{110})^2 \right]
\end{aligned}$$

Using this approximation,

$$\begin{aligned}
E(NSOI) &\cong E_{111} + E_{001} + E_{100} + E_{010} + E_{011} + E_{101} + E_{110} + \frac{(E_{111})(E_{001})(E_{100})(E_{010})}{(E_{011})(E_{101})(E_{110})} \\
&+ \frac{1}{2} \left[\frac{2E_{111}E_{001}E_{100}E_{010}}{E_{011}^3E_{101}E_{110}}V(X_{011}) + \frac{2E_{111}E_{001}E_{100}E_{010}}{E_{101}^3E_{011}E_{110}}V(X_{101}) + \frac{2E_{111}E_{001}E_{100}E_{010}}{E_{110}^3E_{101}E_{011}}V(X_{110}) \right]
\end{aligned}$$

For variance purposes the terms of the approximation involving the second partial derivatives are ignored. Thus we want the variance of the following expression:

$$\begin{aligned}
NSOI &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} \\
&+ \frac{E_{001}E_{100}E_{010}}{E_{011}E_{101}E_{110}}X_{111} + \frac{E_{111}E_{100}E_{010}}{E_{011}E_{101}E_{110}}X_{001} + \frac{E_{111}E_{001}E_{010}}{E_{011}E_{101}E_{110}}X_{100} + \frac{E_{111}E_{001}E_{100}}{E_{011}E_{101}E_{110}}X_{010} \\
&+ \frac{E_{111}E_{001}E_{100}E_{010}}{E_{011}^2E_{101}E_{110}}X_{011} + \frac{E_{111}E_{001}E_{100}E_{010}}{E_{101}^2E_{011}E_{110}}X_{101} + \frac{E_{111}E_{001}E_{100}E_{010}}{E_{110}^2E_{101}E_{011}}X_{110}
\end{aligned}$$

Using obvious notation, write this as

$$\begin{aligned}
NSOI &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} \\
&+ \Delta_{111}X_{111} + \Delta_{001}X_{001} + \Delta_{100}X_{100} + \Delta_{010}X_{010} \\
&+ \Delta_{011}X_{011} + \Delta_{101}X_{101} + \Delta_{110}X_{110}
\end{aligned}$$

And the variance is as follows:

$$\begin{aligned} V(NSOI) &= (1 + \Delta_{111})^2 V(X_{111}) + (1 + \Delta_{001})^2 V(X_{001}) + (1 + \Delta_{100})^2 V(X_{100}) + (1 + \Delta_{010})^2 V(X_{010}) \\ &+ (1 + \Delta_{011})^2 V(X_{011}) + (1 + \Delta_{101})^2 V(X_{101}) + (1 + \Delta_{110})^2 V(X_{110}) \\ &+ (1 + \Delta_{111})(1 + \Delta_{110}) \text{cov}(X_{111}, X_{110}) + (1 + \Delta_{111})(1 + \Delta_{011}) \text{cov}(X_{111}, X_{011}) + (1 + \Delta_{111})(1 + \Delta_{010}) \text{cov}(X_{111}, X_{010}) \\ &+ (1 + \Delta_{110})(1 + \Delta_{011}) \text{cov}(X_{110}, X_{011}) + (1 + \Delta_{110})(1 + \Delta_{010}) \text{cov}(X_{110}, X_{010}) + (1 + \Delta_{011})(1 + \Delta_{010}) \text{cov}(X_{011}, X_{010}) \\ &+ (1 + \Delta_{101})(1 + \Delta_{001}) \text{cov}(X_{101}, X_{001}) \end{aligned}$$

Appendix 2: Expected Value and Variance of Jointly Independent Model

Use the notation from Appendix 1.

The estimator of total population Jl is approximated using Taylor Linearization about the vector of expected values of the seven observed counts resulting in,

$$\begin{aligned}
 Jl &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} + \frac{(E_{001})(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} \\
 &- \frac{E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})^2} (X_{111} - E_{111}) - \frac{E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})^2} (X_{101} - E_{101}) \\
 &- \frac{E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})^2} (X_{011} - E_{011}) + \frac{(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} (X_{001} - E_{001}) \\
 &+ \frac{E_{001}}{(E_{111} + E_{101} + E_{011})} (X_{110} - E_{110}) + \frac{E_{001}}{(E_{111} + E_{101} + E_{011})} (X_{100} - E_{100}) + \frac{E_{001}}{(E_{111} + E_{101} + E_{011})} (X_{010} - E_{010}) \\
 &+ \frac{1}{2} \left[\frac{2E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} (X_{111} - E_{111})^2 + \frac{2E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} (X_{101} - E_{101})^2 + \frac{2E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} (X_{011} - E_{011})^2 \right]
 \end{aligned}$$

Using this approximation,

$$\begin{aligned}
 E(Jl) &\cong E_{111} + E_{001} + E_{100} + E_{010} + E_{011} + E_{101} + E_{110} + \frac{(E_{001})(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} \\
 &+ \frac{1}{2} \left[\frac{2E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} V(X_{111}) + \frac{2E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} V(X_{101}) + \frac{2E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} V(X_{011}) \right]
 \end{aligned}$$

For variance purposes the terms of the approximation involving the second partial derivatives are ignored. Thus we want the variance of the following expression:

$$\begin{aligned}
 Jl &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} \\
 &- \frac{E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})^2} X_{111} - \frac{E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})^2} X_{101} \\
 &- \frac{E_{001}(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})^2} X_{011} + \frac{(E_{110} + E_{100} + E_{010})}{(E_{111} + E_{101} + E_{011})} X_{001} \\
 &+ \frac{E_{001}}{(E_{111} + E_{101} + E_{011})} X_{110} + \frac{E_{001}}{(E_{111} + E_{101} + E_{011})} X_{100} + \frac{E_{001}}{(E_{111} + E_{101} + E_{011})} X_{010}
 \end{aligned}$$

Using obvious notation (same as for NSOI except the partial derivatives are different, write this as

$$Jl \cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110}$$

$$+ \Delta_{111}X_{111} + \Delta_{001}X_{001} + \Delta_{100}X_{100} + \Delta_{010}X_{010}$$

$$+ \Delta_{011}X_{011} + \Delta_{101}X_{101} + \Delta_{110}X_{110}$$

And the variance is as follows:

$$V(Jl) = (1 + \Delta_{111})^2V(X_{111}) + (1 + \Delta_{1001})^2V(X_{001}) + (1 + \Delta_{100})^2V(X_{100}) + (1 + \Delta_{010})^2V(X_{010})$$

$$+ (1 + \Delta_{011})^2V(X_{011}) + (1 + \Delta_{101})^2V(X_{101}) + (1 + \Delta_{110})^2V(X_{110})$$

$$+ (1 + \Delta_{111})(1 + \Delta_{110})\text{cov}(X_{111}, X_{110}) + (1 + \Delta_{111})(1 + \Delta_{011})\text{cov}(X_{111}, X_{011}) + (1 + \Delta_{111})(1 + \Delta_{010})\text{cov}(X_{111}, X_{010})$$

$$+ (1 + \Delta_{110})(1 + \Delta_{011})\text{cov}(X_{110}, X_{011}) + (1 + \Delta_{110})(1 + \Delta_{010})\text{cov}(X_{110}, X_{010}) + (1 + \Delta_{011})(1 + \Delta_{010})\text{cov}(X_{011}, X_{010})$$

$$+ (1 + \Delta_{101})(1 + \Delta_{001})\text{cov}(X_{101}, X_{001})$$

Appendix 3: Expected Value and Variance of Conditionally Independent Model

Use the notation from appendix 1.

The estimator of total population CI is approximated using Taylor Linearization about the vector of expected values of the seven observed counts resulting in,

$$\begin{aligned}
 CI &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} + \frac{(E_{001})(E_{100})}{E_{101}} \\
 &+ \frac{E_{100}}{E_{101}}(X_{001} - E_{001}) + \frac{E_{001}}{E_{101}}(X_{100} - E_{100}) - \frac{E_{001}E_{100}}{E_{101}^2}(X_{101} - E_{101}) \\
 &+ \frac{1}{2} \left[\frac{2E_{001}E_{100}}{E_{101}^3} (X_{101} - E_{101})^2 \right]
 \end{aligned}$$

Using this approximation,

$$\begin{aligned}
 E(CI) &\cong E_{111} + E_{001} + E_{100} + E_{010} + E_{011} + E_{101} + E_{110} + \frac{(E_{001})(E_{100})}{E_{101}} \\
 &+ \frac{1}{2} \left[\frac{2E_{001}E_{100}}{E_{101}^3} V(X_{101}) \right]
 \end{aligned}$$

For variance purposes the terms of the approximation involving the second partial derivatives are ignored. Thus we want the variance of the following expression:

$$\begin{aligned}
 CI &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} \\
 &+ \frac{E_{100}}{E_{101}} X_{001} + \frac{E_{001}}{E_{101}} X_{100} - \frac{E_{001}E_{100}}{E_{101}^2} X_{101}
 \end{aligned}$$

Using obvious notation (same as for TSE except the partial derivatives are different, write this as

$$\begin{aligned}
 CI &\cong X_{111} + X_{001} + X_{100} + X_{010} + X_{011} + X_{101} + X_{110} \\
 &+ \Delta_{001} X_{001} + \Delta_{100} X_{100} + \Delta_{101} X_{101}
 \end{aligned}$$

And the variance is as follows:

$$\begin{aligned}
 V(JI) &= V(X_{111}) + (1 + \Delta_{001})^2 V(X_{001}) + (1 + \Delta_{100})^2 V(X_{100}) + V(X_{010}) \\
 &+ V(X_{011}) + (1 + \Delta_{101})^2 V(X_{101}) + V(X_{110})
 \end{aligned}$$

$$\begin{aligned} &+ \text{cov}(X_{111}, X_{110}) + \text{cov}(X_{111}, X_{011}) + \text{cov}(X_{111}, X_{010}) + (1 + \Delta_{011}) \text{cov}(X_{110}, X_{011}) + \text{cov}(X_{110}, X_{010}) \\ &+ \text{cov}(X_{011}, X_{010}) + (1 + \Delta_{101})(1 + \Delta_{001}) \text{cov}(X_{101}, X_{001}) \end{aligned}$$