# On the Nour (1982) dual-system estimator for population size

Chlebicki Piotr[*], Beręsewicz Maciej[†]

November 23, 2023

**Abstract**

We revisit Nour (1982) [JRSS:A 145(1)] paper on a dual system estimator under positive association between sources. Using Nour's slightly modified algebraic approach to derive an estimate under negatively dependent sources, we find that the expression for the estimate under negative dependence is expressed by the same formula as Nour's proposal. An approximation for the bias and variance of this estimator was derived. The bias of this estimator is non-increasing as a function of the recorded fraction of the population, while maintaining the validity of the assumptions. As an illustration, we present estimates of the population size of large enities with at least one vacancy based solely on online and administrative data for Poland. Our results are comparable with the official statistics published by Statistics Poland in the job vacancy survey.

Keywords: population size estimation, lower bound estimator, capture-recapture, dependent sources, job vacancy survey, dual record system

## Acknowledgements

Piotr Chlebicki is a mathematics master's student at Adam Mickiewicz University in Poznań and his work was conducted in the Educational Research Institute at Warsaw, and funded by the Polish Ministry of Education and Science within a project „Supporting IQS functioning and improvements in order to use its solutions in achieving the country's development strategy aims" (ZSK5). Maciej Beręsewicz was financed by the National Science Center in Poland, OPUS 22 grant no. 2020/39/B/HS4/00941. Piotr Chlebicki is the main author and was responsible for all derivations and calculations in this paper. Maciej Beręsewicz is the corresponding author and he was responsible for the original idea and preparation of the manuscript.

Codes to reproduce the results are freely available on the GitHub repository: https://github.com/ncn-foreigners/paper-nour-note.

[*] piochl@st.amu.edu.pl, Adam Mickiewicz University in Poznań, Faculty of Mathematics and Computer Science, ul. Wieniawskiego 1, 61-712 Poznań

[†] Corresponding author, maciej.beresewicz@ue.poznan.pl, (1) Poznań University of Economics and Business, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań; (2) Statistical Office in Poznań, Poland.

# 1 Introduction

Dependence between sources (often called lists) in capture-recapture studies or multiple system estimation has been, and still is extensively studied. For example, Wolter (1986) who discussed the bivariate Bernoulli models with correlation between random vector explained by time difference, behavioural change or combination of the two and their covariate extensions, Chao et al. (2000) who also considered behavioral and time effects. Nour (1982) considered positive association between lists was considered while negative dependence is also of interest. Recently, Chatterjee and Mukherjee (2018, 2021); Chatterjee and Bhuyan (2020) proposed several estimators for dual system estimation (DSE) for positive and negative dependence.

One exemplary application of estimators in dependent sources in capture-recapture studies are estimates of the scale of modern slavery, based on multiple sources with little overlap in the number of units. (cf. Chan et al., 2021; Sharifi Far et al., 2021). Another example is provided by Beręsewicz et al. (2021), who linked administrative and online data to estimate the number of entities with vacancies and found a slight overlap between these sources. The reason for this small overlap is related to the decision of entities to use specific channels to communicate vacancies.

Table 1: The number of large entities (over 49 employees) with vacancies by sources at the end of 2018Q1

| | Online data ($J$) | | |
|---|---|---|---|
| Admin data ($I$) | Yes ($j = 1$) | No ($j = 0$) | Marginal |
| Yes ($i = 1$) | 534 ($n_{11}$) | 2,584 ($n_{10}$) | 3,118 ($n_{1.}$) |
| No ($i = 0$) | 3,780 ($n_{01}$) | – ($n_{00}$) | ($n_{0.}$) |
| Marginal | 4,314 ($n_{.1}$) | ($n_{.0}$) | $N$ |

Source: Table 7 from Beręsewicz et al. (2021).

Data that motivates our study is presented in Table 1 where $n_{ij}, i, j \in \{0, 1\}$ denotes the number of observations included only: in the first source (register) if $i = 1, j = 0$, in the second source (online) if $i = 0, j = 1$, in both sources if $i = j = 1$ and neither of them if $i = j = 0$. Furthermore let $n_{1.} = n_{11} + n_{10}, n_{.1} = n_{11} + n_{01}$ be the marginal counts. The goal is to estimate unknown number of units not included in both sources (denoted as $n_{00}$) and thus estimate the size of the population (denoted as $N$).

The structure of the paper is as follows. In section 2 we provide basic setup for the derivations and summarise the results of Nour (1982). Section 3 provides derivation under negative dependence. Section 4 provides numerical examples that include simulation study and analysis of the data introduced in the introduction. Paper ends with conclusions and Supplementary materials provide details about the simulation study and derivations of bias and variance.

All calculations were done in the Julia language (Bezanson et al., 2017) and are available at `Github` repository (`https://github.com/ncn-foreigners/paper-nour-note`).

## 2 Basic setup

In this note, we focus on the lower bound population size estimator proposed by Nour (1982) which is based on the estimator for unobserved quantity $n_{00}$ derived for homogeneous (i.e. without individual unit effect) positively correlated lists and under the following assumptions:

**Assumption 1 (A1).** *Positive correlation is present i.e.* $n_{11}n_{00} > n_{10}n_{01}$ *and the dual record system is non-degenerate i.e.* $\forall i, j \in \{0, 1\} : n_{ij} > 0$.

**Assumption 2 (A2).** *Probability that a unit is recorded in either source is greater than* 0.5 *i.e.* $\frac{n_{1\cdot}}{N}, \frac{n_{\cdot 1}}{N} > \frac{1}{2}$. *This assumption ensures that* $n_{11} > n_{00}$.

*In conjunction A1 and A2 guarantee that:* $n_{11}^2 > n_{11}n_{00} > n_{10}n_{01}$.

**Assumption 3 (A3).** *We have:* $n_{10}n_{01} - n_{00}^2 > 0$ *and therefore* $n_{10}n_{01} < \left(\frac{n_{10}n_{01}}{n_{00}}\right)^2$.

Under assumptions $A1, A2, A3$ Nour (1982) derived that the unobserved part of the population $n_{00}$ lies somewhere in the interval[1] $\left[2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01}+n_{11}^2}, \sqrt{n_{10}n_{01}}\right]$ and chose to consider the estimator:

$$\hat{n}_{00} = 2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2}, \tag{1}$$

with the justification that it is most robust with respect to violation of both the third assumption (which in practice is the hardest to verify) and the one we will state later. If the A3 assumption

[1] Assumptions guarantee that $2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2} \leq \sqrt{n_{10}n_{01}}$.

is indeed not violated we are free to choose other estimators such as:

$$\hat{n}_{00} = \sqrt{n_{10}n_{01}}, \tag{2}$$

or even any other point from the $\left(2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01}+n_{11}^2}, \sqrt{n_{10}n_{01}}\right)$ interval.

The $\hat{n}_{00} = \sqrt{n_{10}n_{01}}$ estimator was proposed and discussed in a series of papers of Greenfield (1975, 1976, 1983); Greenfield and Tam (1976) as an *upper bound* estimator for $n_{00}$.

Depending on the selection of estimators (1) or (2) we get a lower and upper bound estimator of the population size under positive dependence denoted as $\hat{N}_L$ and $\hat{N}_U$ respectively:

$$\begin{aligned} \hat{N}_L &= n_{11} + n_{10} + n_{01} + 2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01}+n_{11}^2}, \\ \hat{N}_U &= n_{11} + n_{10} + n_{01} + \sqrt{n_{10}n_{01}}. \end{aligned} \tag{3}$$

## 3 Derivation of estimators for negative dependence system

We noticed that under analogous assumptions the derivation in Nour (1982) can be applied (with slight modification) to the case of negatively correlated sources. In deriving estimates we will modify A1 and A2 to form listed bellow while maintaining the A3 assumption:

**Assumption 1' (A1').** *Negative correlation is present i.e.* $n_{00}n_{11} < n_{10}n_{01}$ *and dual record system is non-degenerate i.e.* $\forall i,j \in \{0,1\} : n_{ij} > 0$.

**Assumption 2' (A2').** *Probability that a unit is recorded in either source is lower than* 0.5 *i.e.* $\frac{n_{1\cdot}}{N}, \frac{n_{\cdot 1}}{N} < \frac{1}{2}$, *where* $N$ *is the true population size. This assumption ensures that* $n_{11} < n_{00}$ *and therefore* $n_{11}^2 < n_{11}n_{00} < n_{01}n_{10}$.

One additional assumption was utilised in Nour (1982), and will also be present in our derivation, that needs to be stated in context of the derivation and hence will be stated while deriving the estimator.

The A1' and A2' assumptions together ensure that for some $K \in \mathbb{R}_+$:

$$\frac{\overbrace{(n_{11} - n_{00})}^{<0}\overbrace{(n_{11}n_{00} - n_{10}n_{01})}^{<0}}{n_{00}n_{1.}n_{.1}} \geq K, \tag{4}$$

$$n_{00}^2 + n_{00}\left(\frac{Kn_{1.}n_{.1} - (n_{11}^2 + n_{10}n_{01})}{n_{11}}\right) + n_{10}n_{01} \leq 0, \tag{5}$$

holds for $n_{00}$ in some possibly degenerate interval $[n_{00}^-, n_{00}^+]$. By Vietta's formulas $n_{00}^-, n_{00}^+$ satisfy equations:

$$n_{00}^- + n_{00}^+ = \frac{n_{11}^2 + n_{10}n_{01} - Kn_{1.}n_{.1}}{n_{11}} \qquad n_{00}^- \cdot n_{00}^+ = n_{01}n_{10}. \tag{6}$$

Applying the geometric-arithmetic mean inequality to (6) yields:

$$\sqrt{n_{01}n_{10}} \leq \frac{n_{11}^2 + n_{10}n_{01} - Kn_{1.}n_{.1}}{2n_{11}} \implies \frac{\left(n_{11} - \sqrt{n_{10}n_{01}}\right)^2}{n_{1.}n_{.1}} \geq K, \tag{7}$$

where the equality in (7) corresponds to: $n_{00}^- = n_{00}^+ = \sqrt{n_{10}n_{01}}$. On the other hand from (6) we have that:

$$n_{00}^+ = \frac{n_{10}n_{01}}{n_{00}^-} \qquad n_{00}^- + \frac{n_{10}n_{01}}{n_{00}^-} < \frac{n_{11}^2 + n_{10}n_{01}}{n_{11}}. \tag{8}$$

Solving this quadratic yields:

$$n_{00}^- \in (n_{11}, \sqrt{n_{10}n_{01}}] \quad n_{00}^+ \in \left[\sqrt{n_{10}n_{01}}, \frac{n_{10}n_{01}}{n_{11}}\right). \tag{9}$$

Now if $n_{00} \in [n_{00}^-, n_{00}^+]$ then there exists $W \in [0, 1]$ satisfying $n_{00} = (1 - W)n_{00}^- + Wn_{00}^+$ which may also be expressed (using (6)) as:

$$W = \frac{n_{00} - n_{00}^-}{n_{00}^+ - n_{00}^-} = \frac{n_{00}n_{00}^- - \left(n_{00}^-\right)^2}{n_{10}n_{01} - \left(n_{00}^-\right)^2}.$$

If we look at $W$ as a function of $n_{00}^{-}$ (for constant values of $n_{00}, n_{10}, n_{01}$) then we have that:

$$\begin{aligned}\frac{dW}{dn_{00}^{-}} &= \frac{\left(n_{00}^{-}\right)^2 n_{00} - 2n_{00}^{-}n_{10}n_{01} + n_{00}n_{10}n_{01}}{\left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)^2} = 0 \\ \iff n_{00}^{-} &= \frac{n_{10}n_{01}}{n_{00}} \pm \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}},\end{aligned} \tag{10}$$

and since $n_{00}^{-} \leq n_{00}^{+}$ we can only consider $n_{00}^{-} = \frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}}$. At this point the second derivative of $W$ is negative since:

$$\begin{aligned}2n_{00}^{-}n_{00} - 2n_{10}n_{01} = -2n_{00}\sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} &< 0 \\ \left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)^2 &> 0 \\ \left(n_{00}^{-}\right)^2 n_{00} - 2n_{00}^{-}n_{10}n_{01} + n_{00}n_{10}n_{01} &= 0 \quad \left\{\text{the numerator of } \frac{dW}{dn_{00}^{-}}\right\},\end{aligned}$$

so it is the maximum of $W$.

Now we state the previously mentioned additional assumption:

**Assumption 4 (A4').** *For a given $n_{00}, n_{10}, n_{01}, n_{11}$ the value of $n_{00}^{-}$ at which maximum of $W$ defined in (10) occurs is contained in the range of possible values for $n_{00}^{-}$ which are given by (9).*

After some elementary manipulations we obtain:

$$n_{11} < n_{00}^{-} = \frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} \implies \frac{2n_{11}n_{10}n_{01}}{n_{11}^2 + n_{10}n_{01}} < n_{00}, \tag{11}$$

and

$$\frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} < \sqrt{n_{10}n_{01}} \implies n_{00} \leq \sqrt{n_{10}n_{01}}. \tag{12}$$

## 3.1 Remarks on the derivation

### 3.1.1 Nour’s third assumption

It should be noted here that the effect of the third assumption (that was again not explicitly stated), which was needed to ensure that $W$ and $\arg\max_{n^-_{00}} W$ are real valued, is equivalent to assuming the upper bound from (12) making this upper bound trivial in our context and even omitted from the entire derivation but we chose to stay near the derivation from original paper.

### 3.1.2 Discrete $W$ function

Secondly it was noted in Macarthur (1983) that in deriving (10) we implicitly assumed that $W$ is a differentiable which is not true. Author derived the actual argument minimum in place of (10) as:

$$n^*_{00} = \frac{n_{10}n_{01}}{n_{00}} - \frac{1}{2} - \sqrt{\frac{1}{4} - n_{10}n_{01} + \left(\frac{n_{10}n_{01}}{n_{00}}\right)^2},$$

and in case of positive correlation between lists results in lower bound estimator in (11) of the form:

$$\hat{n}_{00} = \frac{2n_{11}n_{10}n_{01} + n_{11}^2}{n_{11}^2 + n_{11} + n_{10}n_{01}}, \tag{13}$$

and for negative correlation between lists the resulting estimate is:

$$\hat{n}_{00} = \frac{2n_{11}n_{10}n_{01} + n_{10}n_{01}}{n_{11}^2 + n_{11} + n_{10}n_{01}}. \tag{14}$$

In practice estimates from (13), (14) and (11) do not differ significantly and it is of no consequence which one is used.

### 3.1.3 Formulation in terms of probabilities

The assumptions A1’, A2’ and A3 can be restated using inclusion probabilities for each cell in dual record system as:

**Assumption 1” (A1”, for probabilities).** $\forall i, j \in \{0, 1\} : p_{ij} > 0$ *and* $p_{11}p_{00} > p_{10}p_{01}$

**Assumption 2” (A2”, for probabilities).** $p_{1\cdot}, p_{\cdot 1} > \frac{1}{2}$

**Assumption 3" (A3", for probabilities).** $p_{10}p_{01} > p_{00}^2$

The inequalities (4), (5), (7), (8), (11), (12) can be rewritten as the are by replacing $n$'s by $p$'s and same applies to interval in (9) and equations (6), (10) resulting in the inequality:

$$2\frac{(p_{11}\cdot N)(p_{10}\cdot N)(p_{01}\cdot N)}{(p_{11}\cdot N)^2+(p_{10}\cdot N)(p_{01}\cdot N)} < p_{00}\cdot N < \sqrt{(p_{10}\cdot N)(p_{01}\cdot N)}$$

which gives us the desired estimate for $\hat{n}_{00}$ after substituting full population estimates (which is not observed) for each $p_{ij}\cdot N$.

# 4 Numerical examples

## 4.1 Simulation study

### 4.1.1 Design of the simulation study

In the simulation study use parameterization of the correlated bivariate Bernoulli (BB) distribution which comes from Chatterjee and Bhuyan (2017). Suppose that there exist two uncorrelated variables Bernoulli distributed $I_1^* \sim b(p_1), I_2^* \sim b(p_2)$ with a property that, for variables $I_1, I_2$ which denote the presence or absence in first and second source respectively, the following occurs for the positive (left) and negative (right) dependence:

$$(I_1, I_2)=\begin{cases}(I_1^*, I_1^*) & \text{with prob: } \alpha,\\ (I_1^*, I_2^*) & \text{with prob: } 1-\alpha.\end{cases} \qquad (I_1, I_2)=\begin{cases}(I_1^*, 1-I_1^*) & \text{with prob: } \alpha,\\ (I_1^*, I_2^*) & \text{with prob: } 1-\alpha.\end{cases}$$

in other words $I_1^*, I_2^*$ are equal to $I_1, I_2$ if the process which causes sources to be correlated does not occur and an appropriate modification of $I_1^*, I_2^*$ is the value of $I_1, I_2$ if it does occur. In this parametrization, $\alpha = 0$ corresponds to independence of lists. The probabilities for each cell in $2\times 2$ contingency table (as in table 5) in terms of $p_1, p_2, \alpha$ are given by:

For positive dependence:

$$p_{11} = \alpha p_1 + (1-\alpha)p_1 p_2$$

$$p_{10} = (1-\alpha)p_1(1-p_2)$$

$$p_{01} = (1-\alpha)(1-p_1)p_2$$

$$p_{00} = \alpha(1-p_1) + (1-\alpha)(1-p_1)(1-p_2)$$

$$p_{\cdot 1} = \alpha p_1 + (1-\alpha)p_2$$

$$p_{1\cdot} = p_1,$$

For negative dependence:

$$p_{11} = (1-\alpha)p_1 p_2$$

$$p_{10} = \alpha p_1 + (1-\alpha)p_1(1-p_2)$$

$$p_{01} = \alpha(1-p_1) + (1-\alpha)(1-p_1)p_2$$

$$p_{00} = (1-\alpha)(1-p_1)(1-p_2)$$

$$p_{\cdot 1} = \alpha(1-p_1) + (1-\alpha)p_2$$

$$p_{1\cdot} = p_1.$$

In the simulation study, we consider six scenarios as presented in the Table 2. In the scenarios S1-S4 all assumptions are met while for S5 and S6 assumption A2' is violated. Values for $p_1, p_2$ and $\alpha$ are set by authors and reflects possible coverage of the first and second list and the share of dependent units. Table contains two additional columns: $p_{1.}$ and $p_{2.}$ which refer the coverage of list based on $(p_1, p_2, \alpha)$ hyper-parameters.

Table 2: In the simulation study 4 we considered the following hyper-parameters:

| Scenario | $p_1$ | $p_2$ | $\alpha$ | $p_{1\cdot}$ (Derived) | $p_{\cdot 1}$ (Derived) |
|---|---|---|---|---|---|
| S1 (all assumptions met) | 0.45 | 0.35 | 0.3 | 0.45 | 0.41 |
| S2 (almost independent lists) | 0.45 | 0.45 | 0.005 | 0.45 | 0.4505 |
| S3 (low coverage) | 0.35 | 0.35 | 0.225 | 0.35 | 0.4175 |
| S4 (very low coverage) | 0.15 | 0.15 | 0.05 | 0.15 | 0.185 |
| S5 (slight violation) | 0.5 | 0.5 | 0.2 | 0.5 | 0.5 |
| S6 (substantive violation) | 0.55 | 0.65 | 0.2 | 0.55 | 0.61 |

In the simulation study we test the performance of $\hat{N}_L, \hat{N}_U$ estimators as defined in (3) and

$$\hat{N}_M = \frac{1}{2}\left(\hat{N}_L + \hat{N}_U\right),$$

in an negatively dependent dual record system. We report the simulation results (with $R = 100{,}000$ replicates) using simulation bias ($\frac{1}{R}\sum_{r=1}^{R}\hat{N}_d - N$, where $d \in \{L, U, M\}$) and variance $\left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{N}_d - \bar{\hat{N}}_d\right)^2\right)$. In addition, we report the analytical form of bias and variance for all the estimators.

### 4.1.2 Simulation study results

Table 3 presents simulation results for $N = 1,000$ and various hyper-parameters as defined in table 2. When expected coverage of each list is close to 50% then the bias is small (around 5-6%). However, when coverage decreases, the bias becomes substantial, over 60%. This indicates that Nour's estimator should only be used to indicate a non-trivial lower bound if the researcher suspects that only a small fraction of the population was covered by the study. Furthermore, the analytical bias and variance are very close to the one obtained from the simulation studies, which means that a bias corrected estimators may be applied.

Table 3: Simulation results under negative dependence with $N = 1,000$ and hyper-parameters scenarios. Assumption about coverage of dual record system (A2') is not violated in either of these.

| | Simulation | | Analytical | | |
|---|---|---|---|---|---|
| Estimator | Bias | Var | Bias (Nour's) | Bias | Variance |
| | Scenario 1 | | All assumptions met | | |
| $\hat{N}_L$ | -53.9 | 464.4 | -53.3 | -53.4 | 468.0 |
| $\hat{N}_M$ | 7.4 | 355.4 | – | 7.6 | 355.9 |
| $\hat{N}_U$ | 68.6 | 407.1 | – | 68.6 | 405.4 |
| | Scenario 2 | | Almost independent lists | | |
| $\hat{N}_L$ | -60.7 | 388.2 | -59.7 | -59.9 | 386.6 |
| $\hat{N}_M$ | -57.8 | 400.8 | – | -57.4 | 400.6 |
| $\hat{N}_U$ | -54.9 | 423.0 | – | -55.0 | 423.2 |
| | Scenario 3 | | Low coverage | | |
| $\hat{N}_L$ | -156.8 | 501.4 | -156.3 | -156.4 | 503.7 |
| $\hat{N}_M$ | -98.8 | 410.9 | – | -98.6 | 413.1 |
| $\hat{N}_U$ | -40.7 | 469.1 | – | -40.9 | 470.2 |
| | Scenario 4 | | Very low coverage | | |
| $\hat{N}_L$ | -644.7 | 344.8 | -644.5 | -644.5 | 346.1 |
| $\hat{N}_M$ | -593.2 | 372.0 | – | -593.0 | 373.7 |
| $\hat{N}_U$ | -541.6 | 466.6 | – | -541.6 | 466.4 |

Table 4: Simulation results under violation of the A2' assumption.

|  | Simulation |  | Analytical |  |  |
|---|---|---|---|---|---|
| Estimator | Bias | Var | Bias (Nour's) | Bias | Variance |
|  | Scenario 5 |  | Slight violation |  |  |
| $\hat{N}_L$ | 75.8 | 302.3 | 76.9 | 76.7 | 299.2 |
| $\hat{N}_M$ | 87.7 | 300.7 | – | 88.2 | 298.6 |
| $\hat{N}_U$ | 99.7 | 341.6 | – | 99.8 | 340.0 |
|  | Scenario 6 |  | More substantive violation |  |  |
| $\hat{N}_L$ | 165.4 | 240.3 | 166.4 | 166.1 | 238.9 |
| $\hat{N}_M$ | 165.8 | 242.8 | – | 166.2 | 241.5 |
| $\hat{N}_U$ | 166.3 | 246.1 | – | 166.2 | 245.3 |

Table 4 contains simulation results for cases when second assumption is violated. In all cases, estimates were higher than $N = 1,000$. Indicating that second assumption is indeed crucial. Figure 1 presents contour plots of the relative bias of $\hat{N}_L$ depending on the parameters $p_1, p_2$ and $\alpha$. Plots indicate that the bias becomes significant as $p_1$ and $p_2$ depart from 0.5.

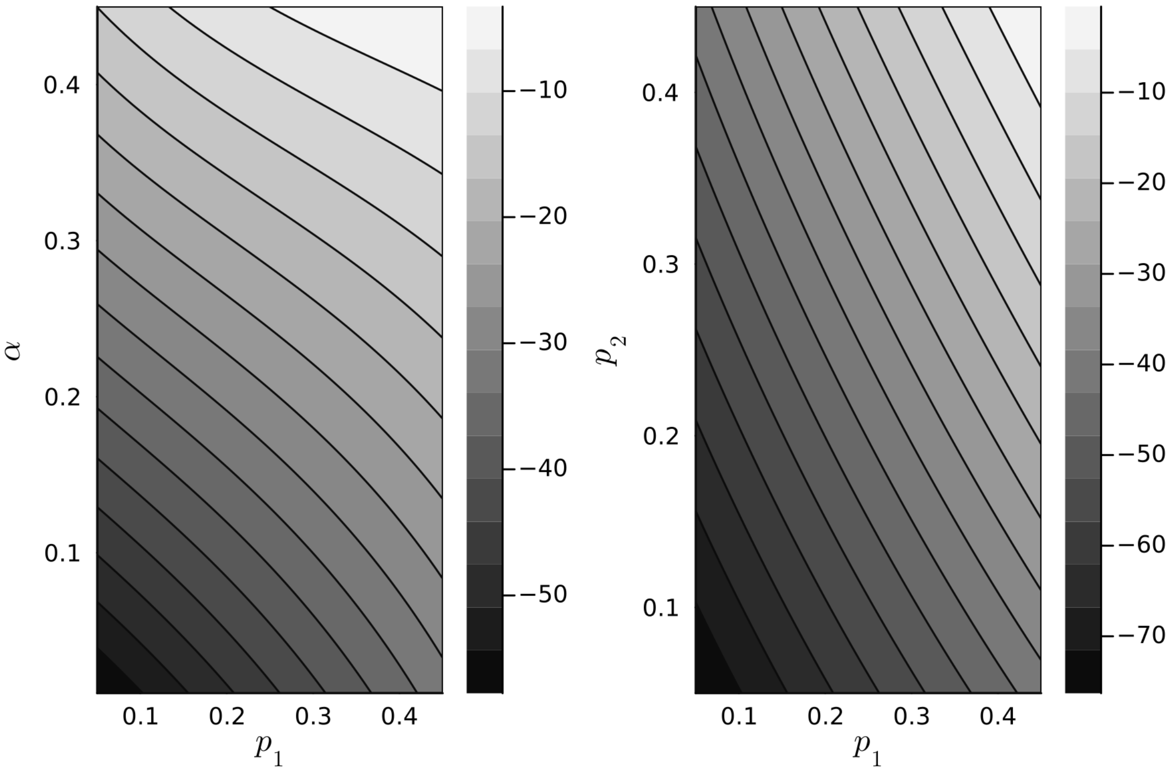

Figure 1: Contour plot of relative bias of $\hat{N}_L$ when $p_2$ is fixed at 0.35 (left) or $\alpha$ is fixed at 0.15 (right)

## 4.2 Empirical study

Finally, we present results for the motivation example presented in table 1. We expect low coverage between sources due to decisions by large entities to use selected channels to reach potential employees. According to the Job Vacancy Survey (JVS) for this quarter, about 7.3 thousand large enterprises (with more than 49 employees) had vacancies (cf. Beręsewicz et al. (2021, p. 36, table 8)). This suggests that administrative data and online data cover about 58% and 42% of large datasets respectively. Note that for online data the suggested coverage is higher than the assumed 50%, but this proportion is calculated based on the point estimate reported by the JVS and we do not know the confidence interval for the estimated number of units with vacancies.

Table 5: Estimated number of large entities (over 49 employees) with vacancies by sources at the end of 2018Q1

| $n$ | $\hat{n}_{00,L}$ | $\hat{n}_{00,U}$ | $\hat{N}_L$ | $\widehat{Var}(\hat{N}_L)$ | $\hat{N}_U$ | $\widehat{Var}(\hat{N}_U)$ | $\hat{N}_{\text{naive}}$ |
|---|---|---|---|---|---|---|---|
| 6,898 | 1,038 | 3,125 | 7,936 | 2,832 | 10,023 | 4,485 | 25,189 |

According to the presented estimators, the lower bound estimator of $n_{00}$ denoted as $\hat{n}_{00,L}$ under negative dependence (11) is about 1k while the upper bound estimator (12) denoted as $\hat{n}_{00,U}$ is over 3k entities. This suggests that the possible number of large units with vacancies is between 8k ($\hat{N}_L$) and 10k ($\hat{N}_U$), which is larger than reported in the JVS survey. For these data, the naive dual-system estimator denoted by $\hat{N}_{\text{naive}}$ is significantly larger, suggesting that over 25k large enterprises have at least one vacancy, which would represent about 40% of all large enterprises in the population (about 65k).

# 5 Summary

In this note we showed that Nour's lower bound estimator of the population size under negative dependence has the same formula as for positive dependence. We derived bias and variance approximations for the lower and upper bounds of the unobserved part $n_{00}$ and its average. We showed that the bias of Nour's estimator increases as the coverage of the lists decreases from $\frac{1}{2}$, but the property that the estimated population size is reliably lower than the true population size is maintained.

In the simulation study, we have shown that the assumed coverage of the lists used in the study (i.e. the A2' assumption) is the most critical for this estimator. This suggests that using the lower and upper bounds of Nour's estimator may be useful if we assume that the coverage of each source is close to 50% of the study population. While the lower bound may prove usefull if researchers are interested in obtaining non trivial lower bound on $N$ even if coverage is much lower than 50%.

Finally, the empirical study addressed the problem of estimating the number of units with vacancies based solely on administrative and online data. The results presented in the paper are in line with the official data from the JVS survey, which suggest that the application of Nour's estimator under negative dependence is the right choice for this problem.

The paper and the results presented in it may be useful for researchers and practitioners interested in estimating population size based on non-statistical data sources. Online and administrative data can be an interesting alternative for the population rarely observed in official statistics and measured by surveys.

# References

Beręsewicz, M., H. Cherniaiev, and R. Pater (2021). Estimating the number of entities with vacancies using administrative and online data.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2017). Julia: A fresh approach to numerical computing. *SIAM Review 59*(1), 65–98.

Chan, L., B. W. Silverman, and K. Vincent (2021). Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists. *Journal of the American Statistical Association 116*(535), 1297–1306.

Chao, A., W. Chu, and C.-H. Hsu (2000). Capture–recapture when time and behavioral response affect capture probabilities. *Biometrics 56*(2), 427–433.

Chatterjee, K. and P. Bhuyan (2017). On the estimation of population size from a post-stratified two-sample capture–recapture data under dependence. *Journal of Statistical Computation and Simulation 90*, 819 – 838.

Chatterjee, K. and P. Bhuyan (2020). On the estimation of population size from a post-stratified two-sample capture–recapture data under dependence. *Journal of Statistical Computation and Simulation 90*(5), 819–838.

Chatterjee, K. and D. Mukherjee (2018). A new integrated likelihood for estimating population size in dependent dual-record system. *Canadian Journal of Statistics 46*(4), 577–592.

Chatterjee, K. and D. Mukherjee (2021). On the estimation of population size under dependent dual-record system: an adjusted profile-likelihood approach. *Journal of Statistical Computation and Simulation 91*(13), 2740–2763.

Greenfield, C. C. (1975). On the Estimation of a Missing Cell in a 2 × 2 Contingency Table. *Journal of the Royal Statistical Society. Series A (General) 138*(1), 51.

Greenfield, C. C. (1976). A Revised Procedure for Dual Record Systems in Estimating Vital Events. *Journal of the Royal Statistical Society. Series A (General) 139*(3), 389.

Greenfield, C. C. (1983). On Estimators for Dual Record Systems. *Journal of the Royal Statistical Society. Series A (General) 146*(3), 273.

Greenfield, C. C. and S. M. Tam (1976). A Simple Approximation for the Upper Limit to the Value of a Missing Cell in 2 × 2 Contingency Table. *Journal of the Royal Statistical Society. Series A (General) 139*(1), 96.

Macarthur, E. W. (1983). A note on the estimation of vital events: Total number and proportion. *Journal of the Royal Statistical Society: Series A (General) 146*(1), 85–86.

Nour, E.-S. (1982). On the estimation of the total number of vital events with data from dual collection systems. *Journal of the Royal Statistical Society: Series A (General) 145*(1), 106–116.

Sharifi Far, S., R. King, S. Bird, A. Overstall, H. Worthington, and N. Jewell (2021). Multiple systems estimation for modern slavery: Robustness of list omission and combination. *Crime & Delinquency 67*(13-14), 2213–2236.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association 81*(394), 337–346.

# A Supplementary material

## A.1 Derivation of bias and standard error

The second order Taylor series expansion of function of three random variables $Y_1, Y_2, Y_3$ around point $\mathbb{E}(Y_1, Y_2, Y_3) = (\mu_1, \mu_2, \mu_3)$ is given by[2]:

$$\begin{aligned}F(Y_1, Y_2, Y_3) \approx F(\mu_1, \mu_2, \mu_3) &+ \sum_{k=1}^{3} \frac{\partial F}{\partial Y_k}(\mu_1, \mu_2, \mu_3)(Y_k - \mu_k) \\ &+ \frac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{\partial^2 F}{\partial Y_j \partial Y_k}(\mu_1, \mu_2, \mu_3)(Y_k - \mu_k)(Y_j - \mu_j)\end{aligned}$$

It leads to approximation for expected value of $F$ as:

$$\begin{aligned}\mathbb{E}[F(Y_1, Y_2, Y_3)] \approx F(\mu_1, \mu_2, \mu_3) &+ \sum_{k=1}^{3} \frac{\partial F}{\partial Y_k}(\mu_1, \mu_2, \mu_3) \underbrace{\mathbb{E}(Y_k - \mu_k)}_{=0} \\ &+ \frac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{\partial^2 F}{\partial Y_j \partial Y_k}(\mu_1, \mu_2, \mu_3) \mathbb{E}\left[(Y_k - \mu_k)(Y_j - \mu_j)\right] \\ &= F(\mu_1, \mu_2, \mu_3) + \frac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{\partial^2 F}{\partial Y_j \partial Y_k}(\mu_1, \mu_2, \mu_3) \sigma_{Y_k Y_j}\end{aligned} \tag{15}$$

If we restrict ourselves to the first order expansion and apply the var operator to both sides of approximation we get

$$\begin{aligned}\operatorname{var}\left(F(Y_1, Y_2, Y_3)\right) &\approx \operatorname{var}\left( F(\mu_1, \mu_2, \mu_3) + \sum_{k=1}^{3} \frac{\partial F}{\partial Y_k}(\mu_1, \mu_2, \mu_3)(Y_k - \mu_k) \right) \\ &= \sum_{j=1}^{3} \sum_{k=1}^{3} \left( \frac{\partial F}{\partial Y_j} \cdot \frac{\partial F}{\partial Y_k} \right) (\mu_1, \mu_2, \mu_3) \sigma_{Y_k Y_j}\end{aligned} \tag{16}$$

To estimate the expected value and variance of estimators derived above covariance between vari-

[2] Assuming that $F$ is twice differentiable.

ables $n_{11}, n_{01}, n_{10}$ is needed which is given by:

$$\operatorname{cov}(n_{11}, n_{10}, n_{01}) = N \begin{pmatrix} p_{11}(1-p_{11}) & -p_{11}p_{10} & -p_{11}p_{01} \\ -p_{11}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{01} \\ -p_{11}p_{01} & -p_{10}p_{01} & p_{01}(1-p_{01}) \end{pmatrix}$$

Lastly to apply equations (15) (16) to our estimators we need first and second partial derivatives with respect to $n_{11}, n_{10}, n_{01}$. If we set the following notation for derived estimators:

$$\begin{aligned} \hat{N}_{(\mathrm{L})} &= n_{11} + n_{10} + n_{01} + 2\frac{n_{11}n_{10}n_{01}}{n_{11}^2 + n_{10}n_{01}} \\ \hat{N}_{(\mathrm{M})} &= n_{11} + n_{10} + n_{01} + \frac{1}{2}\left(\sqrt{n_{10}n_{01}} + 2\frac{n_{11}n_{10}n_{01}}{n_{11}^2 + n_{10}n_{01}}\right) \\ \hat{N}_{(\mathrm{U})} &= n_{11} + n_{10} + n_{01} + \sqrt{n_{10}n_{01}} \end{aligned} \tag{17}$$

Of course these are defined only for discrete values of $n_{11}, n_{10}, n_{01} \in \mathbb{N}^3$ but functions in (17) are just discretizations of appropriate functions from $\mathbb{R}_+^3$ to $\mathbb{R}_+$. If we apply (restricted) Taylor series expansion to these functions at values $n_{11}, n_{10}, n_{01}$ we also get a valid approximations for values of (17) (since these functions coincide on $\mathbb{N}^3$). Therefore we use the $\dfrac{\partial N_{\mathrm{U}}}{\partial n_{11}}$ etc. as a shorthand for derivatives of these appropriate differentiable functions. The derivatives are as follows:

- For $\hat{N}_{(\mathrm{U})}$

$$\begin{aligned} \frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{11}} &= 1 \\ \frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{01}} &= 1 + \frac{n_{10}}{2\sqrt{n_{10}n_{01}}} = 1 + \frac{1}{2}\sqrt{\frac{n_{10}}{n_{01}}} \\ \frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{10}} &= 1 + \frac{n_{01}}{2\sqrt{n_{10}n_{01}}} = 1 + \frac{1}{2}\sqrt{\frac{n_{01}}{n_{10}}} \\ \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{11}\partial n_{01}} = \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{11}\partial n_{10}} &= \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{11}^2} = 0 \\ \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{10}\partial n_{01}} = \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{01}\partial n_{10}} &= \frac{1}{4}\frac{1}{\sqrt{n_{10}n_{01}}} \\ \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{01}^2} &= -\frac{1}{4}\frac{\sqrt{n_{10}}}{\sqrt{n_{01}^3}} \\ \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{10}^2} &= -\frac{1}{4}\frac{\sqrt{n_{01}}}{\sqrt{n_{10}^3}} \end{aligned} \tag{18}$$

- For $\hat{N}_{(\mathrm{L})}$

$$\begin{aligned}
\frac{\partial \hat{N}_{(\mathrm{L})}}{\partial n_{11}} &= 1 + 2\frac{n_{10}n_{01}\left(n_{11}^2 + n_{10}n_{01}\right) - 2n_{11}^2 n_{10}n_{01}}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} \\
&= 1 + 2\frac{n_{10}^2 n_{01}^2 - n_{11}n_{10}n_{01}}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} = \frac{n_{11}^4 + 3n_{10}^2 n_{01}^2}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} \\
\frac{\partial \hat{N}_{(\mathrm{L})}}{\partial n_{01}} &= 1 + 2\frac{n_{11}n_{10}\left(n_{11}^2 + n_{10}n_{01}\right) - n_{10}^2 n_{11}n_{01}}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} \\
&= 1 + 2\frac{n_{10}n_{11}^3}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} \\
\frac{\partial \hat{N}_{(\mathrm{L})}}{\partial n_{10}} &= 1 + 2\frac{n_{11}n_{01}\left(n_{11}^2 + n_{10}n_{01}\right) - n_{10}n_{11}n_{01}^2}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} \\
&= 1 + 2\frac{n_{01}n_{11}^3}{\left(n_{11}^2 + n_{10}n_{01}\right)^2} \\
\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{11}\partial n_{01}} &= \frac{6n_{10}^2 n_{01}\left(n_{11}^2 + n_{10}n_{01}\right)^2}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
&- \frac{2\left(n_{11}^2 + n_{10}n_{01}\right)n_{10}\left(n_{11}^4 + 3n_{10}^2 n_{01}^2\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{11}\partial n_{10}} &= \frac{6n_{10}n_{01}^2\left(n_{11}^2 + n_{10}n_{01}\right)^2}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
&- \frac{2\left(n_{11}^2 + n_{10}n_{01}\right)n_{01}\left(n_{11}^4 + 3n_{10}^2 n_{01}^2\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4}
\end{aligned} \tag{19}$$

$$\begin{aligned}
\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{01}\partial n_{11}} &= 2\frac{3n_{11}^2 n_{10}\left(n_{11}^2 + n_{10}n_{01}\right)^2 - 4n_{11}^4 n_{10}\left(n_{11}^2 + n_{10}n_{01}\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{10}\partial n_{11}} &= 2\frac{3n_{11}^2 n_{01}\left(n_{11}^2 + n_{10}n_{01}\right)^2 - 4n_{11}^4 n_{01}\left(n_{11}^2 + n_{10}n_{01}\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{11}^2} &= 4\frac{n_{11}n_{10}n_{01}\left(n_{11}^4 - 2n_{11}^2 n_{10}n_{01} - 3n_{10}^2 n_{01}^2\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
&= 4\frac{n_{11}n_{10}n_{01}\left(n_{11}^2 - 3n_{10}n_{01}\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4}
\end{aligned} \tag{20}$$

$$\begin{aligned}
\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{10}\partial n_{01}} &= \frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{01}\partial n_{10}} = 2\frac{n_{11}^3\left(n_{11}^2 + n_{10}n_{01}\right)\left(n_{11}^2 - n_{10}n_{01}\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \\
&= 2\frac{n_{11}^3\left(n_{11}^2 - n_{10}n_{01}\right)}{\left(n_{11}^2 + n_{10}n_{01}\right)^3}
\end{aligned} \tag{21}$$

$$\frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{01}^2} = -4\frac{n_{11}^3 n_{10}^2}{\left(n_{11}^2 + n_{10}n_{01}\right)^4} \quad \frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{10}^2} = -4\frac{n_{11}^3 n_{01}^2}{\left(n_{11}^2 + n_{10}n_{01}\right)^4}$$

- For $\hat{N}_{(\mathrm{M})}$ in general we have that:

$$\hat{N}_{(\mathrm{M})} = \frac{1}{2}\left(\hat{N}_{(\mathrm{U})} + \hat{N}_{(\mathrm{L})}\right)$$

so the derivatives of this estimator are expressed by the ones above with adjustment being:

$$\begin{aligned}\frac{\partial \hat{N}_{(\mathrm{M})}}{\partial n_{ij}} &= \frac{1}{2}\left(\frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{ij}} + \frac{\partial \hat{N}_{(\mathrm{L})}}{\partial n_{ij}}\right)\\ \frac{\partial^2 \hat{N}_{(\mathrm{M})}}{\partial n_{i_1j_1}\partial n_{i_2j_2}} &= \frac{1}{2}\left(\frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{i_1j_1}\partial n_{i_2j_2}} + \frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{i_1j_1}\partial n_{i_2j_2}}\right)\end{aligned} \tag{22}$$

We may now substitute the results of (18) (19) (22) and covariance matrix into the approximations (15) and (16) giving us:

- For estimator $\hat{N}_{\mathrm{U}}$:

$$\begin{aligned}\mathbb{E}[\hat{N}_{\mathrm{U}}] \approx{}& N\left(p_{11} + p_{10} + p_{01} + \sqrt{p_{10}p_{01}}\right)\\ &- \frac{1}{2}\left(\frac{\sqrt{p_{10}p_{01}}}{2} + \frac{(1-p_{10})}{4}\sqrt{\frac{p_{01}}{p_{10}}} + \frac{(1-p_{01})}{4}\sqrt{\frac{p_{10}}{p_{01}}}\right)\\ \mathrm{var}\left(\hat{N}_{\mathrm{U}}\right) ={}& N\left(p_{11}(1-p_{11}) + p_{10}(1-p_{10}) + p_{01}(1-p_{01})\right)\\ &+ Np_{10}(1-p_{10})\left(\sqrt{\frac{p_{01}}{p_{10}}} + \frac{1}{4}\frac{p_{01}}{p_{10}}\right)\\ &+ Np_{01}(1-p_{01})\left(\sqrt{\frac{p_{10}}{p_{01}}} + \frac{1}{4}\frac{p_{10}}{p_{01}}\right)\\ &- 2Np_{11}p_{01}\left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) - 2Np_{11}p_{10}\left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right)\\ &- 2Np_{10}p_{01}\left(1 + \frac{1}{4} + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}} + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right)\end{aligned} \tag{23}$$

- For estimator $\hat{N}_{\mathrm{L}}$:

$$
\begin{aligned}
\mathbb{E}[\hat{N}_{\mathrm{L}}] \approx\; & N\left(p_{11}+p_{10}+p_{01}+2\frac{p_{11}p_{10}p_{01}}{p_{11}^2+p_{10}p_{01}}\right)\\
& + 2\frac{p_{11}(1-p_{11})}{N^2}p_{11}p_{10}p_{01}\frac{p_{11}^2-3p_{10}p_{01}}{\left(p_{11}^2+p_{10}p_{01}\right)^4}\\
& - 2\frac{p_{10}(1-p_{10})}{N^2}\frac{p_{11}^3p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^4}\\
& - 2\frac{p_{01}(1-p_{01})}{N^2}\frac{p_{11}^3p_{10}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^4}\\
& - 2p_{11}p_{10}\frac{3p_{11}^2p_{01}\left(p_{11}^2+p_{10}p_{01}\right)^2-4p_{11}^4p_{01}\left(p_{11}^2+p_{10}p_{01}\right)}{\left(p_{11}^2+p_{10}p_{01}\right)^4}\\
& - 2p_{11}p_{01}\frac{3p_{11}^2p_{10}\left(p_{11}^2+p_{10}p_{01}\right)^2-4p_{11}^4p_{10}\left(p_{11}^2+p_{10}p_{01}\right)}{\left(p_{11}^2+p_{10}p_{01}\right)^4}\\
& - 2p_{10}p_{01}\frac{p_{11}^3\left(p_{11}^3-p_{10}p_{01}\right)}{\left(p_{11}^2+p_{10}p_{01}\right)^3}\\
\operatorname{var}\left(\hat{N}_{\mathrm{L}}\right) \approx\; & Np_{11}(1-p_{11})\left(\frac{p_{11}^4+3p_{10}^2p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)^2 \qquad (24)\\
& + Np_{01}(1-p_{01})\left(1+2\frac{p_{10}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)^2\\
& + Np_{10}(1-p_{10})\left(1+2\frac{p_{01}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)^2\\
& - 2Np_{10}p_{01}\left(1+2\frac{p_{01}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)\\
& \cdot\left(1+2\frac{p_{10}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)\\
& - 2Np_{11}p_{10}\left(1+2\frac{p_{01}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)\\
& \cdot\left(\frac{p_{11}^4+3p_{10}^2p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)\\
& - 2Np_{11}p_{01}\left(1+2\frac{p_{10}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)\\
& \cdot\left(\frac{p_{11}^4+3p_{10}^2p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)
\end{aligned}
$$

- For estimator $\hat{N}_{\mathrm{M}}$ the expected value will be equal to the mean of $\hat{N}_L$ and $\hat{N}_U$, and variance will be expressed as:

$$
\begin{aligned}
\mathbb{E}[\hat{N}_{\mathrm{M}}] &= \frac{\mathbb{E}[\hat{N}_{\mathrm{L}}] + \mathbb{E}[\hat{N}_{\mathrm{U}}]}{2} \\
\operatorname{var}\left(\hat{N}_{\mathrm{M}}\right) &= \frac{1}{4}\left(\operatorname{var}\left(\hat{N}_{\mathrm{L}}\right) + \operatorname{var}\left(\hat{N}_{\mathrm{U}}\right)\right) \\
&\quad + \frac{N}{4}\Bigg(2p_{11}(1-p_{11})\left(\frac{p_{11}^4 + 3p_{10}^2p_{01}^2}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad + 2p_{01}(1-p_{01})\left(1 + 2\frac{p_{10}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot\left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right) \\
&\quad + 2p_{10}(1-p_{10})\left(1 + 2\frac{p_{01}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot\left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) \\
&\quad - 2p_{10}p_{01}\Bigg(\left(1 + 2\frac{p_{01}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot\left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) + \left(1 + 2\frac{p_{10}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot\left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right)\Bigg) \\
&\quad - 2p_{10}p_{11}\Bigg(\left(\frac{p_{11}^4 + 3p_{10}^2p_{01}^2}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) \\
&\quad + \left(1 + 2\frac{p_{01}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right)\Bigg)\Bigg) \\
&\quad - 2p_{11}p_{01}\Bigg(\left(\frac{p_{11}^4 + 3p_{10}^2p_{01}^2}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right) \\
&\quad + \left(1 + 2\frac{p_{10}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right)\Bigg)\Bigg)\Bigg)
\end{aligned}
\tag{25}
$$

which was achieved by exploiting the property:

$$\operatorname{var}\left(\hat{N}_{\mathrm{M}}\right) \approx \sum_{\substack{j\in\{01,\\10,11\}}} \sum_{\substack{k\in\{01,\\10,11\}}} \left(\frac{\partial \hat{N}_{\mathrm{M}}}{\partial n_j} \cdot \frac{\partial \hat{N}_{\mathrm{M}}}{\partial n_k}\right)(Np_{11}, Np_{10}, Np_{01})\sigma_{Y_kY_j} =$$

$$\frac{1}{4} \sum_{\substack{j\in\{01,\\10,11\}}} \sum_{\substack{k\in\{01,\\10,11\}}} \left(\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_k} + \frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_k} + \frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_k} + \frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_k}\right)(Np_{11}, Np_{10}, Np_{01})\sigma_{Y_kY_j} =$$

$$\frac{1}{4}\left(\operatorname{var}\left(\hat{N}_{\mathrm{L}}\right) + \operatorname{var}\left(\hat{N}_{\mathrm{U}}\right)\right) + \frac{1}{4} \sum_{\substack{j\in\{01,\\10,11\}}} \sum_{\substack{k\in\{01,\\10,11\}}} \left(\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_k} + \frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_k}\right)(Np_{11}, Np_{10}, Np_{01})\sigma_{Y_kY_j}$$