A note on Nour (1982) dual-system estimator for negative dependence between sources

Chlebicki Piotr*, Beręsewicz Maciej†

May 11, 2023

Abstract

We revisit Nour (1982) [JRSS:A 145(1)] paper on a dual system estimator under positive dependency between sources. We show that the original lower bound estimator is the same under negative dependence between sources when the coverage of each source is at most 50% of the population. We derived a slightly better approximation of bias in Nour's estimator taking into account dependence between sources. We show that the bias of this estimator increases as the coverage of lists decreases.

Keywords: population size estimation, lower bound estimator, capture-recapture, dependent sources.

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 is the main author. He was responsible for all derivations and calculations in this paper. Maciej is the corresponding author. 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 in capture-recapture studies has been studied by Wolter (1986) who discussed the M_b model or Chao et al. (2000) who considered behavioral and time effects. Recently, Chatterjee and Mukherjee (2018, 2021); Chatterjee and Bhuyan (2020) proposed several estimators for dual system estimation (DSE) for positive and negative dependence.

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 positive dependence between lists and under the following assumptions :

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. $n_{00}, n_{\cdot 1}, n_{1\cdot} > 0$, where n_{ij} denotes the number of observations in the list i -th and j -th where $i, j \in \{0, 1\}$ where 1 indicate presence in the list i, j and 0 otherwise. $n_{\cdot 1}, n_{1\cdot}$ denote marginal frequencies.

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:

$$n_{11}^2 > n_{11}n_{00} > n_{10}n_{01}$$

A3 It is assumed that

$$n_{10}n_{01} - n_{00}^2 > 0 \tag{1}$$

and thus

$$n_{10}n_{01} < \left(\frac{n_{10}n_{01}}{n_{00}}\right)^2$$

Unit inclusions are only correlated through the dependence between sources. Although this assumption was not stated by Nour (1982), it can be derived from 1-2 assumptions. One additional assumption will be outlined when it is used due to the specific nature of why the need for that specific assumption arises (cf. Supplementary A eq (16)).

Nour (1982) chose to consider the estimator

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

with the justification that it may be robust to the breaking of the third assumption.

If the third assumption is not violated we are free to choose other estimators such as

$$\hat{n}_{00} = \sqrt{n_{10}n_{01}}, \tag{3}$$

or even

$$\hat{n}_{00} = \frac{1}{2}\left(2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2} + \sqrt{n_{10}n_{01}}\right), \tag{4}$$

which have not been considered in Nour (1982). It should be noted that $\hat{n}_{00} = \sqrt{n_{10}n_{01}}$ 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 (2) or (3) 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

$$\begin{aligned} \hat{N}_L &= n_{11} + n_{10} + n_{01} + 2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2}, \\ \hat{N}_U &= n_{11} + n_{10} + n_{01} + \sqrt{n_{10}n_{01}}. \end{aligned} \tag{5}$$

Nour (1982) considered only positive dependence between lists while in practice we may observe negative dependence. For example, Beręsewicz et al. (2021) linked administrative and online data to estimate the number of entities with vacancies and detected a slight overlap between these sources which is presented in Table 1. Other cases may refer to the estimation of modern slavery where overlap between lists is very scarce (cf. Chan et al., 2021; Sharifi Far et al., 2021).

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

	Online data	
Admin data	Yes	No
Yes	534	2,584
No	3,780	–

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

Although both data sources may cover a significant part of the study population, the overlap is small, which makes the estimates based on the naive dual-system estimator (Lincoln-Petersen) very large (over 25k for example in Table 1 which is over 3 times higher than the official estimates provided by Statistics Poland). Therefore, this motivates this work on Nour's estimator for negative dependence.

The structure of the paper is as follows. In section 2 we rewrite Nour (1982) derivation under positive dependence using standard dual-system estimation notation which will be used in the next section. In section 3 we derive estimators for negative dependence and show that they are equivalent to those presented in the introduction and section 2. Additionally, we derive slight bias and variance for these estimators. Section 4 provides simulation study results as well as results for the motivating example from Table 1.

2 Derivation of estimators for positive dependence system

The first two assumptions above ensure that the quantity, which lies at approach in finding estimator for n_{00} in method presented by Nour:

$$\frac{(n_{11} - n_{00})(n_{11}n_{00} - n_{10}n_{01})}{n_{00}n_{1\cdot}n_{\cdot 1}} \tag{6}$$

is non-negative and thus there exists $0 < K \in \mathbb{R}$ such that a quadratic inequality (with respect to n_{00}) occurs:

$$\frac{(n_{11} - n_{00})(n_{11}n_{00} - n_{10}n_{01})}{n_{00}n_{1\cdot}n_{\cdot 1}} \geq K \tag{7}$$

which may also be rewritten by multiplying by positive quantities $n_{00}, n_{\cdot 1}, n_{1\cdot}$ rearranging the terms and dividing by $-n_{11}$ as:

$$n_{00}^2 + n_{00}\left(\frac{Kn_{1\cdot}n_{\cdot 1} - (n_{11}^2 + n_{10}n_{01})}{n_{11}}\right) + n_{10}n_{01} \leq 0 \tag{8}$$

has solutions in an interval[1] $[n_{00}^-, n_{00}^+]$ n_{00}^-, n_{00}^+ being roots of quadratic polynomial $n_{00}^- \leq n_{00}^+$.

From Vieta's formulas for quadratic polynomial we see that the following relations between observed quantities and roots of polynomial in (8):

$$\begin{aligned} n_{00}^- + n_{00}^+ &= \frac{n_{11}^2 + n_{10}n_{01} - Kn_{1\cdot}n_{\cdot 1}}{n_{11}} \\ n_{00}^- \cdot n_{00}^+ &= n_{01}n_{10} \end{aligned} \tag{9}$$

which leads us to bound for K if we apply the geometric mean and arithmetic mean inequality:

$$\begin{aligned} \sqrt{n_{01}n_{10}} &\leq \frac{n_{11}^2 + n_{10}n_{01} - Kn_{1\cdot}n_{\cdot 1}}{2n_{11}} \\ \implies \frac{\left(n_{11} - \sqrt{n_{10}n_{01}}\right)^2}{n_{1\cdot}n_{\cdot 1}} &\geq K \end{aligned} \tag{10}$$

the upper bound from (10) corresponds to: $n_{00}^- = n_{00}^+ = \sqrt{n_{10}n_{01}}$ and the other bound on K that is $K = 0$ gives us:

$$\begin{aligned} n_{00}^+ &= \frac{n_{10}n_{01}}{n_{00}^-} \quad \text{from (4)} \\ 0 &= \left(n_{00}^+\right)^2 - n_{00}^+\left(n_{11} + \frac{n_{10}n_{01}}{n_{11}}\right) + n_{10}n_{01} \end{aligned} \tag{11}$$

which finally leads to (remembering that we put $n_{00}^- < n_{00}^+$):

$$n_{00}^- = \frac{n_{10}n_{01}}{n_{11}}, n_{00}^+ = n_{11}$$

The reasoning above proves that:

$$n_{00}^- \in \left[\frac{n_{10}n_{01}}{n_{11}}, \sqrt{n_{10}n_{01}}\right) \quad n_{00}^+ \in (\sqrt{n_{10}n_{01}}, n_{11}] \tag{12}$$

[1]The interval may be degenerate if instead of inequalities $n_{11}^2 > n_{11}n_{00} > n_{10}n_{01}$ the equality $n_{11} = n_{00} = \sqrt{n_{10}n_{01}}$ holds.

since increase in quantity K corresponds to decrease in n_{00}^- and increase in n_{00}^+.

The statement $n_{00} \in \left[n_{00}^-, n_{00}^+\right]$ implies that there exists $W \in [0,1]$ such that:

$$n_{00} = W n_{00}^+ + (1-W) n_{00}^- \tag{13}$$

which may be described explicitly as (second equation follows from substituting for n_{00}^+ from (9)):

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

If we look at W as a function of n_{00}^- (for constant values of n_{00}, n_{10}, n_{01}) it's derivative will be given as:

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

$$\frac{dW}{dn_{00}^-} = 0 \overset{2}{\iff} n_{00}^- = \frac{n_{10} n_{01}}{n_{00}} \pm \sqrt{\left(\frac{n_{10} n_{01}}{n_{00}}\right)^2 - n_{10} n_{01}} \tag{14}$$

From (9) and $n_{00}^- < n_{00} < n_{00}^+$ we have that:

$$n_{00}^- = \frac{n_{10} n_{01}}{n_{00}^+} < \frac{n_{10} n_{01}}{n_{00}}$$

which is only satisfied by one root in (14) that is

$$n_{00}^- = \frac{n_{10} n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10} n_{01}}{n_{00}}\right)^2 - n_{10} n_{01}} \tag{15}$$

furthermore the second derivative of W is given by:

$$\begin{aligned}
\frac{d^2 W}{d\left(n_{00}^-\right)^2} = \Bigg(& \left(2 n_{00}^- n_{00} - 2 n_{10} n_{01}\right) \left(n_{10} n_{01} - \left(n_{00}^-\right)^2\right)^2 \\
& + 4 n_{00}^- \left(n_{10} n_{01} - \left(n_{00}^-\right)^2\right) \left(\left(n_{00}^-\right)^2 n_{00} - 2 n_{00}^- n_{10} n_{01} + n_{00} n_{10} n_{01}\right) \Bigg) \\
& \Big/ \left(n_{10} n_{01} - \left(n_{00}^-\right)^2\right)^4
\end{aligned}$$

[2] Here the third assumption was utilised, although it was not explicitly stated in Nour (1982).

we have that at (15):

$$\begin{aligned} 2n_{00}^{-}n_{00} - 2n_{10}n_{01} = -2n_{00}\sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} < 0 \\ \left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)^2 > 0 \\ n_{10}n_{01} - \left(n_{00}^{-}\right)^2 > 0 \quad \text{from (12)} \\ \left(n_{00}^{-}\right)^2 n_{00} - 2n_{00}^{-}n_{10}n_{01} + n_{00}n_{10}n_{01} = 0 \end{aligned}$$

where the last equation follows simply from substitution of (15) to the numerator of $\frac{d^2W}{d\left(n_{00}^{-}\right)^2}$.

These relations imply that

$$\frac{d^2W}{d\left(n_{00}^{-}\right)^2} < 0$$

at (15) which means that (15) is the point that corresponds to local maximum for W.

Now to finally derive the estimate for $\hat{n}_{00}$ we need a third assumption:

For a given $n_{00}, n_{10}, n_{01}, n_{11}$ the value of n_{00}^{-} at which maximum of W occurs described by equation (15) has the same range of possible values as n_{00}^{-} which are given by (12).

If that assumption holds we have that

$$\begin{aligned} &\frac{n_{10}n_{01}}{n_{11}} \leq \frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} \\ &\implies n_{00} \geq 2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2} \end{aligned} \tag{16}$$

$$\begin{aligned} &\frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} < \sqrt{n_{10}n_{01}} \\ &\implies n_{00} < \sqrt{n_{10}n_{01}} \end{aligned} \tag{17}$$

both obtained by rearranging terms and squaring both sides of the inequalities. Conjunction of (16) and (17) gives us the full bound:

$$2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2} \leq n_{00} < \sqrt{n_{10}n_{01}} \tag{18}$$

which gives us a few possible choices of estimators. Nour (1982) chose to consider estimator

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

with the justification that it is more robust with respect to the violation of third assumption. If that assumption is not in fact violated we're free to choose other estimators such as

$$\hat{n} = \sqrt{n_{10}n_{01}}$$

or even

$$\hat{n} = \frac{1}{2}\left(2\frac{n_{10}n_{01}n_{11}}{n_{10}n_{01} + n_{11}^2} + \sqrt{n_{10}n_{01}}\right)$$

which have not been considered in Nour (1982). It will be shown later in numeric simulation that in fact the lower bound estimator is in fact worse than $\hat{n} = \sqrt{n_{10}n_{01}}$. Then, as stated in the introduction lower and upper bounds for n_{00} can be used to derive lower and upper bound estimators of the population size given by (5). Bias and variance approximations are given in Supplementary Materials in (31)-(33).

Remark on the derivation

It should be noted here that the effect of the third assumption (that was again not explicitly stated), which was needed to ensure that (14) and W are real valued, is equivalent to assuming the upper bound from (18) 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 Derivation of estimators for negative dependence system

If we were to extend the previous approach for negative dependence, the assumptions listed below would be analogous to the ones in Nour (1982):

- Negative correlation is present i.e.

$$n_{00}n_{11} < n_{10}n_{01}$$

  and dual record system is non-degenerate i.e. $n_{00}, n_{10}, n_{01}, n_{11} > 0$

- Probability that a unit is recorded in either source is less than 0.5 i.e.

$$\frac{n_{1\cdot}}{N}, \frac{n_{\cdot 1}}{N} < \frac{1}{2}$$

  we have that

$$n_{11} < n_{00} \implies n_{11}^2 < n_{11}n_{00} < n_{01}n_{10}$$

- We have that $n_{00}^2 < n_{10}n_{01}$ which will again be used to allow the existence of root $\frac{dW}{dn_{00}^-} = 0$.

- The unit inclusions are only correlated through the correlation between sources. Although this assumption was not stated it was present.

The first two assumptions together ensure that:

$$\underbrace{(n_{11} - n_{00})}_{\leq 0}\underbrace{(n_{11}n_{00} - n_{10}n_{01})}_{\leq 0} \Big/ n_{00}n_{1\cdot}n_{\cdot 1}$$

is positive and so the inequality from (8) stands for some $K \in \mathbb{R}_{>0}$ and has solutions is some $[n_{00}^-, n_{00}^+]$ that also satisfy equations (9) obtained using Vieta's formula. The bounds from (12) are now modified to:

$$n_{00}^- \in [n_{11}, \sqrt{n_{10}n_{01}}) \quad n_{00}^+ \in \left(\sqrt{n_{10}n_{01}}, \frac{n_{10}n_{01}}{n_{11}}\right] \tag{19}$$

since in the case of positive dependence we have:

$$n_{11}^2 > n_{11}n_{00} > n_{10}n_{01}$$

but now we have:

$$n_{11}^2 < n_{11}n_{00} < n_{10}n_{01}$$

from which it follows that the roots of a quadratic:

$$0 = \left(n_{00}^{+}\right)^2 - n_{00}^{+}\left(n_{11} + \frac{n_{10}n_{01}}{n_{11}}\right) + n_{10}n_{01}$$

now have opposite order with respect to $<$ albeit the $n_{11} - n_{10}n_{01}/n_{11}$ term is now negative so we ought to take its additive inverse for the square root of the quadratic discriminant, so (12) still stands, the problem arises in the derivative of W:

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

we have that the discriminant of the numerator is:

$$4n_{10}n_{01}\left(n_{10}n_{01} - n_{00}^2\right) \tag{20}$$

and it is not always positive. If fact simple numerical experiments show that even in case of positive dependence and when assumptions put forth by Nour are met that term is not always positive and that is where the third assumption comes.

In case of negative dependence, we assumed that probabilities of being included in either source are less than $\frac{1}{2}$, for the third assumption to hold these probabilities need to be 'close' to $\frac{1}{2}$. For instance, when p_1 and p_2 are set to $0.3, 0.284$ respectively, the term above was never positive in the simulation on the other hand, with values $0.45, 0.47$ it was positive in 99.737% of cases.

If the 3rd assumption has met the root of $\frac{dW}{dn_{00}^{-}}$ in (14) remains unchanged and the same argument gives us that (15) as the value of n_{00}^{-} at which this root occurs.

Moreover at (15) the second derivative of W given by:

$$\begin{aligned}\frac{d^2W}{d\left(n_{00}^{-}\right)^2} = \Bigg(& \left(2n_{00}^{-}n_{00} - 2n_{10}n_{01}\right)\left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)^2 \\ & + 4n_{00}^{-}\left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)\left(\left(n_{00}^{-}\right)^2 n_{00} - 2n_{00}^{-}n_{10}n_{01} + n_{00}n_{10}n_{01}\right)\Bigg) \\ & \Big/ \left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)^4\end{aligned}$$

and has the same sign as in the case of positive dependence:

$$
\begin{aligned}
2n_{00}^{-}n_{00} - 2n_{10}n_{01} = -2n_{00}\sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} < 0 \\
\left(n_{10}n_{01} - \left(n_{00}^{-}\right)^2\right)^2 > 0 \\
n_{10}n_{01} - \left(n_{00}^{-}\right)^2 > 0 \quad \text{from (19)} \\
\left(n_{00}^{-}\right)^2 n_{00} - 2n_{00}^{-}n_{10}n_{01} + n_{00}n_{10}n_{01} = 0 \quad \text{numerator of } \frac{dW}{dn_{00}^{-}}
\end{aligned}
$$

Now once again we need one more assumption to find estimators of n_{00}:

For a given $n_{00}, n_{10}, n_{01}, n_{11}$ the value of n_{00}^{-} at which maximum of W occurs described by equation (15) has the same range of possible values as n_{00}^{-} which are given by (19),

which leads us to:

$$
\begin{aligned}
n_{11} &\leq \frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} \\
\implies \frac{2n_{11}n_{10}n_{01}}{n_{11}^2 + n_{10}n_{01}} &\leq n_{00}
\end{aligned}
\tag{21}
$$

and

$$
\begin{aligned}
0 < \frac{n_{10}n_{01}}{n_{00}} - \sqrt{\left(\frac{n_{10}n_{01}}{n_{00}}\right)^2 - n_{10}n_{01}} &< \sqrt{n_{10}n_{01}} \\
\implies n_{00} &< \sqrt{n_{10}n_{01}}
\end{aligned}
\tag{22}
$$

(22) also follows directly from 4th assumption.

Since properties of $\hat{N}_{\mathrm{U}}, \hat{N}_{\mathrm{M}}, \hat{N}_{\mathrm{L}}$ derived in (31), (32), (33) only depended on the form of estimators and were derived for arbitrary $p_{11}, p_{01}, p_{10}, p_{00}$ they still hold in case of negative dependence. Macarthur (1983) remarks about the lack of continuity in W and estimating the standard error of the estimates also remain true and relevant.

4 Numerical examples

In the numerical examples, we assess assumptions and the quality of estimators. We use the following parameterisation of the correlated bivariate Bernoulli (BB) distribution which comes from Chatterjee and Bhuyan (2017).

Suppose that there exist two uncorrelated variables Bernoulli distributed $I_1^* \sim \text{Bern}(p_1), I_2^* \sim \text{Bern}(p_2)$ 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:

$$(I_1, I_2) \overset{3}{=} \begin{cases} (I_1^*, I_1^*) & \text{with prob: } \alpha, \\ (I_1^*, I_2^*) & \text{with prob: } 1-\alpha. \end{cases} \qquad (I_1, I_2) \overset{4}{=} \begin{cases} (I_1^*, 1-I_1^*) & \text{with prob: } \alpha, \\ (I_1^*, I_2^*) & \text{with prob: } 1-\alpha. \end{cases}$$

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. The probabilities for each cell in 2x2 contingency table in terms of p_1, p_2, α are given by:

For positive association:	For negative association:
$p_{11} = \alpha p_1 + (1-\alpha)p_1 p_2$	$p_{11} = (1-\alpha)p_1 p_2$
$p_{10} = (1-\alpha)p_1(1-p_2)$	$p_{10} = \alpha p_1 + (1-\alpha)p_1(1-p_2)$
$p_{01} = (1-\alpha)(1-p_1)p_2$	$p_{01} = \alpha(1-p_1) + (1-\alpha)(1-p_1)p_2$
$p_{00} = \alpha(1-p_1) + (1-\alpha)(1-p_1)(1-p_2)$	$p_{00} = (1-\alpha)(1-p_1)(1-p_2)$
$p_{\cdot 1} = \alpha p_1 + (1-\alpha)p_2$	$p_{\cdot 1} = \alpha(1-p_1) + (1-\alpha)p_2$
$p_{1\cdot} = p_1,$	$p_{1\cdot} = p_1.$

4.1 Assumption's assessment

We perform numerical integration to determine the proportion of parameter space in which the assumptions are met by probabilities the check how often the following assumptions hold for both negative as well as positive associations between sources:

$(A1)$ $p_{1\cdot}, p_{\cdot 1} > \frac{1}{2}$ for positive dependence and $p_{1\cdot}, p_{\cdot 1} < \frac{1}{2}$ for negative

$(A2)$ $\left(\frac{p_{10}p_{01}}{p_{00}}\right)^2 > p_{10}p_{01}$

[1]For the positive association between sources.
[4]For the negative association between sources.

($A3$) The assumptions about the range of

$$\arg\max_{n_{00}^{-}} W$$

cannot be easily verified numerically instead we verify whether two previous assumptions hold in conjunction with $p_{00} \leq 2\frac{p_{11}p_{10}p_{01}}{p_{11}^2+p_{10}p_{01}}$ which would be implied by this assumption holding true. This is in essence a conclusion that N_L is a lower bound estimate.

Results of the verification of assumptions are presented in Table 2. For the negative case, when we consider the whole parameter's space the A1 assumption is met only 12% times, while for positive around 37%. The second assumption (A2) is met 13% times for positive and over 83% times for negative dependence between sources.

Table 2: Numerical verification[6] of the assumptions for the positive and negative dependence between sources

Dependence	$A1$	$A2$	$A3$	$A2 \mid A1$	$A3 \mid A1$	$A3 \mid A2$	$A3 \mid A1 \wedge A2$
Positive	0.3744	0.1310	0.9581	0.2655	0.9298	0.6798	0.7357
Negative	0.1245	0.8360	0.4036	0.3202	0.9685	0.2866	0.9017

Source: own elaboration. We set α, p_1, and p_2 range from 0.005 to 0.995 by 0.0025.

It should be noted that if the first assumption is met, the third assumption about the p_{00} is met over 92% times for positive and over 96% for negative cases. This indicates that even if most assumptions made by Nour are indeed violated but coverage of both lists is as demanded the Nour's 'lower bound estimator' could be used to reliably indicate the lower bound of the population size, which could be considered a more convincing justification for Nour's estimate than his algebraic manipulations.

[6] It should be noted that reparametrization would change the results obtained. For example in an extreme case if we used parametrisation with $\phi = \frac{p_{00}p_{11}}{p_{01}p_{10}}$ instead of α to capture dependence between sourced the Lebesgue measure of parameter space would be infinite.

4.2 Quality of estimators

In this section, we present the results of the simulation study for negative dependence for $\hat{N}_L, \hat{N}_M$ and $\hat{N}_U$ estimators as defined in (5) when first assumption (the easiest to verify and most crucial) is met or violated. We report the simulation results (with 100 000 replicates) as well as the analytical form of bias and variance for all the estimators discussed.

Table 3 presents simulation results for 4 cases with various values of p_1, p_2 and α and $N = 1,000$. When lists' coverage is close to the assumed levels, i.e. 50% then the bias is small (around 5-6%). However, when coverage decreases, the bias becomes substantial, close to 60%. This indicates that this estimator should not be used or only used to indicate a non-trivial lower bound if the researcher assumes that input lists cover only a small fraction of the population 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 various values of p_1, p_2 and α so the first assumption holds.

	Simulation		Analytical		
Estimator	Bias	Var	Bias (Nour's)	Bias	Variance
			$p_1 = 0.45, p_2 = 0.35, \alpha = 0.30$		
$\hat{N}_L$	-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
			$p_1 = 0.45, p_2 = 0.45, \alpha = 0.005$		
$\hat{N}_L$	-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
			$p_1 = 0.35, p_2 = 0.35, \alpha = 0.225$		
$\hat{N}_L$	-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
			$p_1 = 0.15, p_2 = 0.15, \alpha = 0.05$		
$\hat{N}_L$	-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 contains simulation results for two cases when the coverage of the lists is over 50%

and negative dependence is assumed. In all cases, estimates are higher than $N = 1,000$ which is indicated by the positive bias.

Table 4: Simulation results under negative dependence with $N = 1000$ and various values of p_1, p_2 and α, so the first assumption is violated

	Simulation		Analytical		
Estimator	Bias	Var	Bias (Nour's)	Bias	Variance
		$p_1 = 0.50, p_2 = 0.50, \alpha = 0.20$			
$\hat{N}_L$	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
		$p_1 = 0.55, p_2 = 0.65, \alpha = 0.30$			
$\hat{N}_L$	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

Figure 1 presents contour plots of the relative bias of $\hat{N}_L$ depending on the parameters p_1, p_2 and α. Plots indicate that the bias becomes significant as p_1 and p_2 depart from 0.5.

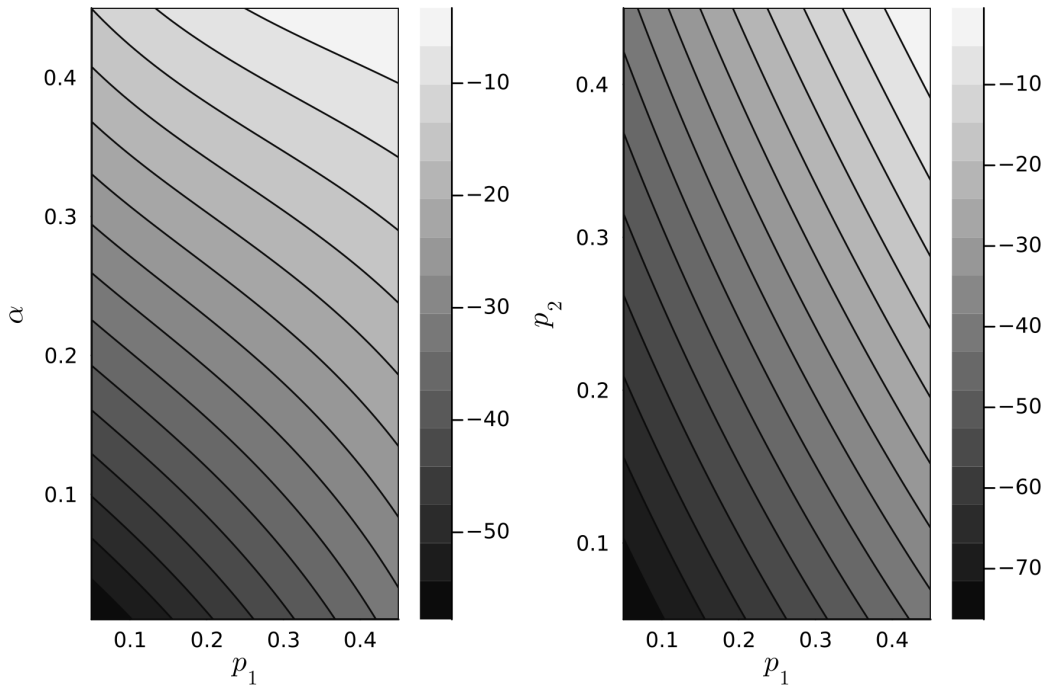

Figure 1: Contour plot of relative bias of $\hat{N}_L$ when p_2 is fixed at 0.35 (left) or α is fixed at 0.15 (right)

Finally, we present results for the motivating example presented in Table 1. According to discussed estimators, the lower bound of n_{00} under negative dependence is about 1k while the

upper is over 3k entities. This suggests that the possible number of large entities with vacancies may be between 8k ($\hat{N}_L$) and 10k ($\hat{N}_U$), which is close to the estimates based on Statistics Poland's Job Vacancy Survey for this quarter (7,262; cf. Beręsewicz et al. (2021, p. 36, table 8)).

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

5 Summary

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

References

Beręsewicz, M., H. Cherniaiev, and R. Pater (2021). Estimating the number of entities with vacancies using administrative and online data. *arXiv preprint arXiv:2106.03263*.

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

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[7]:

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

it leads to approximation for expected value of F as:

$$\begin{aligned} \mathbb{E}[F(Y_1, Y_2, Y_3)] &\approx F(\mu_1, \mu_2, \mu_3) + \sum_{k=1}^{3} \frac{\partial F}{\partial Y_k}(\mu_1, \mu_2, \mu_3) \underbrace{\mathbb{E}(Y_k - \mu_k)}_{=0} \\ &+ \frac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{\partial^2 F}{\partial Y_j \partial Y_k}(\mu_1, \mu_2, \mu_3) \mathbb{E}\left[(Y_k - \mu_k)(Y_j - \mu_j)\right] \\ &= F(\mu_1, \mu_2, \mu_3) + \frac{1}{2} \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{\partial^2 F}{\partial Y_j \partial Y_k}(\mu_1, \mu_2, \mu_3) \sigma_{Y_k Y_j} \end{aligned} \tag{23}$$

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

$$\begin{aligned} \operatorname{var}\left(F(Y_1, Y_2, Y_3)\right) &\approx \operatorname{var}\left(F(\mu_1, \mu_2, \mu_3) + \sum_{k=1}^{3} \frac{\partial F}{\partial Y_k}(\mu_1, \mu_2, \mu_3)(Y_k - \mu_k) \right) \\ &= \sum_{j=1}^{3} \sum_{k=1}^{3} \left(\frac{\partial F}{\partial Y_j} \cdot \frac{\partial F}{\partial Y_k} \right) (\mu_1, \mu_2, \mu_3) \sigma_{Y_k Y_j} \end{aligned} \tag{24}$$

To estimate the expected value and variance of estimators derived above covariance between variables n_{11}, n_{01}, n_{10} is needed which is given by:

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

[7] Assuming that F is twice differentiable.

Lastly to apply equations (23) (24) 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 estimators deriving from inequalities (18):

$$\begin{aligned}
\hat{N}_{(\mathrm{L})} &= n_{11} + n_{10} + n_{01} + 2\frac{n_{11}n_{10}n_{01}}{n_{11}^2 + n_{10}n_{01}} \\
\hat{N}_{(\mathrm{M})} &= n_{11} + n_{10} + n_{01} + \frac{1}{2}\left(\sqrt{n_{10}n_{01}} + 2\frac{n_{11}n_{10}n_{01}}{n_{11}^2 + n_{10}n_{01}}\right) \\
\hat{N}_{(\mathrm{U})} &= n_{11} + n_{10} + n_{01} + \sqrt{n_{10}n_{01}}
\end{aligned} \tag{25}$$

From which it follows that their derivatives are given by:

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

$$\begin{aligned}
\frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{11}} &= 1 \\
\frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{01}} &= 1 + \frac{n_{10}}{2\sqrt{n_{10}n_{01}}} = 1 + \frac{1}{2}\sqrt{\frac{n_{10}}{n_{01}}} \\
\frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{10}} &= 1 + \frac{n_{01}}{2\sqrt{n_{10}n_{01}}} = 1 + \frac{1}{2}\sqrt{\frac{n_{01}}{n_{10}}} \\
\frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{11}\partial n_{01}} &= \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{11}\partial n_{10}} = \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{11}^2} = 0 \\
\frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{10}\partial n_{01}} &= \frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{01}\partial n_{10}} = \frac{1}{4}\frac{1}{\sqrt{n_{10}n_{01}}} \\
\frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{01}^2} &= -\frac{1}{4}\frac{\sqrt{n_{10}}}{\sqrt{n_{01}^3}} \\
\frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{10}^2} &= -\frac{1}{4}\frac{\sqrt{n_{01}}}{\sqrt{n_{10}^3}}
\end{aligned} \tag{26}$$

- 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} \\
&\quad - \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} \\
&\quad - \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} \qquad (27) \\
\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} \qquad (28) \\
\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} \qquad (29) \\
\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}
\end{aligned}
$$

- 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:

$$\begin{aligned}
\frac{\partial \hat{N}_{(\mathrm{M})}}{\partial n_{ij}} &= \frac{1}{2}\left(\frac{\partial \hat{N}_{(\mathrm{U})}}{\partial n_{ij}} + \frac{\partial \hat{N}_{(\mathrm{L})}}{\partial n_{ij}}\right) \\
\frac{\partial^2 \hat{N}_{(\mathrm{M})}}{\partial n_{i_1 j_1}\partial n_{i_2 j_2}} &= \frac{1}{2}\left(\frac{\partial^2 \hat{N}_{(\mathrm{U})}}{\partial n_{i_1 j_1}\partial n_{i_2 j_2}} + \frac{\partial^2 \hat{N}_{(\mathrm{L})}}{\partial n_{i_1 j_1}\partial n_{i_2 j_2}}\right)
\end{aligned} \tag{30}$$

We may now substitute the results of (26) (27) (30) and covariance matrix into the approximations (23) and (24) giving us:

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

$$\begin{aligned}
\mathbb{E}[\hat{N}_{\mathrm{U}}] \approx\ & N\left(p_{11} + p_{10} + p_{01} + \sqrt{p_{10}p_{01}}\right) \\
& - \frac{1}{2}\left(\frac{\sqrt{p_{10}p_{01}}}{2} + \frac{(1-p_{10})}{4}\sqrt{\frac{p_{01}}{p_{10}}} + \frac{(1-p_{01})}{4}\sqrt{\frac{p_{10}}{p_{01}}}\right) \\
\mathrm{var}\left(\hat{N}_{\mathrm{U}}\right) =\ & N\left(p_{11}(1-p_{11}) + p_{10}(1-p_{10}) + p_{01}(1-p_{01})\right) \\
& + Np_{10}(1-p_{10})\left(\sqrt{\frac{p_{01}}{p_{10}}} + \frac{1}{4}\frac{p_{01}}{p_{10}}\right) \\
& + Np_{01}(1-p_{01})\left(\sqrt{\frac{p_{10}}{p_{01}}} + \frac{1}{4}\frac{p_{10}}{p_{01}}\right) \\
& - 2Np_{11}p_{01}\left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) - 2Np_{11}p_{10}\left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right) \\
& - 2Np_{10}p_{01}\left(1 + \frac{1}{4} + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}} + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right)
\end{aligned} \tag{31}$$

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

$$
\begin{aligned}
\mathbb{E}[\hat{N}_{\mathrm{L}}] \approx\; & N\left(p_{11}+p_{10}+p_{01}+2\frac{p_{11}p_{10}p_{01}}{p_{11}^2+p_{10}p_{01}}\right) \\
& + 2\frac{p_{11}(1-p_{11})}{N^2}p_{11}p_{10}p_{01}\frac{p_{11}^2-3p_{10}p_{01}}{\left(p_{11}^2+p_{10}p_{01}\right)^4} \\
& - 2\frac{p_{10}(1-p_{10})}{N^2}\frac{p_{11}^3p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^4} \\
& - 2\frac{p_{01}(1-p_{01})}{N^2}\frac{p_{11}^3p_{10}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^4} \\
& - 2p_{11}p_{10}\frac{3p_{11}^2p_{01}\left(p_{11}^2+p_{10}p_{01}\right)^2-4p_{11}^4p_{01}\left(p_{11}^2+p_{10}p_{01}\right)}{\left(p_{11}^2+p_{10}p_{01}\right)^4} \\
& - 2p_{11}p_{01}\frac{3p_{11}^2p_{10}\left(p_{11}^2+p_{10}p_{01}\right)^2-4p_{11}^4p_{10}\left(p_{11}^2+p_{10}p_{01}\right)}{\left(p_{11}^2+p_{10}p_{01}\right)^4} \\
& - 2p_{10}p_{01}\frac{p_{11}^3\left(p_{11}^3-p_{10}p_{01}\right)}{\left(p_{11}^2+p_{10}p_{01}\right)^3} \\
\operatorname{var}\left(\hat{N}_{\mathrm{L}}\right) \approx\; & Np_{11}(1-p_{11})\left(\frac{p_{11}^4+3p_{10}^2p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)^2 \\
& + Np_{01}(1-p_{01})\left(1+2\frac{p_{10}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)^2 \\
& + Np_{10}(1-p_{10})\left(1+2\frac{p_{01}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)^2 \\
& - 2Np_{10}p_{01}\left(1+2\frac{p_{01}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right) \\
& \cdot\left(1+2\frac{p_{10}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right) \\
& - 2Np_{11}p_{10}\left(1+2\frac{p_{01}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right) \\
& \cdot\left(\frac{p_{11}^4+3p_{10}^2p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right) \\
& - 2Np_{11}p_{01}\left(1+2\frac{p_{10}p_{11}^3}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right) \\
& \cdot\left(\frac{p_{11}^4+3p_{10}^2p_{01}^2}{\left(p_{11}^2+p_{10}p_{01}\right)^2}\right)
\end{aligned}
\tag{32}
$$

- 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:

$$
\begin{aligned}
\mathbb{E}[\hat{N}_{\mathrm{M}}] &= \frac{\mathbb{E}[\hat{N}_{\mathrm{L}}] + \mathbb{E}[\hat{N}_{\mathrm{U}}]}{2} \\
\mathrm{var}\left(\hat{N}_{\mathrm{M}}\right) &= \frac{1}{4}\left(\mathrm{var}\left(\hat{N}_{\mathrm{L}}\right) + \mathrm{var}\left(\hat{N}_{\mathrm{U}}\right)\right) \\
&\quad + \frac{N}{4}\Bigg(2p_{11}(1-p_{11})\left(\frac{p_{11}^4 + 3p_{10}^2p_{01}^2}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad + 2p_{01}(1-p_{01})\left(1 + 2\frac{p_{10}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right) \\
&\quad + 2p_{10}(1-p_{10})\left(1 + 2\frac{p_{01}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) \\
&\quad - 2p_{10}p_{01}\Bigg(\left(1 + 2\frac{p_{01}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) + \left(1 + 2\frac{p_{10}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \\
&\quad \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right)\Bigg) \\
&\quad - 2p_{10}p_{11}\Bigg(\left(\frac{p_{11}^4 + 3p_{10}^2p_{01}^2}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{10}}{p_{01}}}\right) \\
&\quad + \left(1 + 2\frac{p_{01}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right)\Bigg) \\
&\quad - 2p_{11}p_{01}\Bigg(\left(\frac{p_{11}^4 + 3p_{10}^2p_{01}^2}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right) \cdot \left(1 + \frac{1}{2}\sqrt{\frac{p_{01}}{p_{10}}}\right) \\
&\quad + \left(1 + 2\frac{p_{10}p_{11}^3}{\left(p_{11}^2 + p_{10}p_{01}\right)^2}\right)\Bigg)\Bigg)
\end{aligned}
\tag{33}
$$

which was achieved by exploiting the property:

$$
\begin{aligned}
\operatorname{var}\left(\hat{N}_{\mathrm{M}}\right) &\approx \sum_{\substack{j\in\{01,\\ 10,11\}}} \sum_{\substack{k\in\{01,\\ 10,11\}}} \left(\frac{\partial \hat{N}_{\mathrm{M}}}{\partial n_j} \cdot \frac{\partial \hat{N}_{\mathrm{M}}}{\partial n_k}\right) (Np_{11}, Np_{10}, Np_{01})\sigma_{Y_k Y_j} \\
&= \frac{1}{4} \sum_{\substack{j\in\{01,\\ 10,11\}}} \sum_{\substack{k\in\{01,\\ 10,11\}}} \left(\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_k} + \frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_k} + \frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_k}\right. \\
&\quad \left. + \frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_k}\right) (Np_{11}, Np_{10}, Np_{01})\sigma_{Y_k Y_j} \\
&= \frac{1}{4}\left(\operatorname{var}\left(\hat{N}_{\mathrm{L}}\right) + \operatorname{var}\left(\hat{N}_{\mathrm{U}}\right)\right) + \frac{1}{4} \sum_{\substack{j\in\{01,\\ 10,11\}}} \sum_{\substack{k\in\{01,\\ 10,11\}}} \left(\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_k}\right. \\
&\quad \left. + \frac{\partial \hat{N}_{\mathrm{L}}}{\partial n_j}\frac{\partial \hat{N}_{\mathrm{U}}}{\partial n_k}\right) (Np_{11}, Np_{10}, Np_{01})\sigma_{Y_k Y_j}
\end{aligned}
$$