Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Nonparametric methods for population size estimation

Presented by: Manjari Das

Accepted by: Department of Statistics & Data Science

Readers:

Edward H. Kennedy, Advisor

Larry Wasserman

Robin Mejia

Sivaraman Balakrishnan

Nicholas P. Jewell, London School of Hygiene and Tropical Medicine

Approved by the Committee on Graduate Degrees:

Richard Scheines, Dean Date

Carnegie Mellon University

Nonparametric methods for population size estimation

A Dissertation Submitted to the Graduate School
in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy
in
Statistics
by

Manjari Das

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

May 2022

c

© by Manjari Das, 2022
All Rights Reserved.

Dedicated to ma, baba, dadu, mammam and Souvik.

Acknowledgements

This work would not have been complete without the contributions and support of many individuals. First, I would like to thank my advisor Professor Edward H. Kennedy who has been an amazing guide to me since 2018. He introduced me to the whole area of capture-recapture and causal inference. He has also helped me in writing my first paper, encouraged me to go to my first big conference in the USA, and has always been there to discuss old and new topics which helped in my growth. What I admire most is him inspiring me to do things myself no matter how many tries it takes while helping me improve after each try. I am especially grateful for his patience and guidance during the last two years amid the pandemic while also managing other important obligations.

I would also like to thank our collaborator Professor Nicholas P. Jewell for the very interesting questions he raised and for blessing us with his expertise during our meetings. I am very grateful to our committee members for their contributions, introducing new ideas and sharing very useful perspectives on our real data analysis. I would like to express my gratitude to Professor Robin for sharing her grad school journey during the dinners she hosted, Professor Larry Wasserman for being there on my committee for both the advanced data analysis and the PhD thesis despite his very busy schedule, and Professor Sivaraman Balakrishnan for being a point of contact after I first received the acceptance letter from CMU. Professor Valerie Ventura for pushing me towards improving my skills in applied statistics.

I want to thank my fellow PhD students Benjamin LeRoy, Ciaran Evans, Boyan Duan, Niccoló Dalmasso, Alan Mishler, Matteo Bonvini, Ian Waudby-Smith, Jisu Kim, Ilmun Kim and my entire cohort for all the help and support at various stages and for making this journey smoother.

I want to acknowledge the role of various teachers throughout my life. My parents and grandparents were my first teachers and instilled an interest in mathematics in me. My high school teacher Mr Sambit Bhowmik for making mathematics fun and guiding me towards cracking my first mathematics olympiad. My mentor Professor Diganta Mukherjee at the Indian Statistical Institute for being my advisor in my first research project and guiding me in pursuing a PhD.

Last, but very importantly, I want to thank my mother, my father and my partner Souvik for being there throughout the ups and downs and for the very much needed constant upliftment in this long journey. My parents have always supported me in my decisions and helped in any way they could, while also guiding me in the right direction. Besides my advisor and other mentors, Souvik has been there helping me in preparing for presentations and interviews for the last decade. I am fortunate to have them in my life.

“Being a successful person is not necessarily defined by what you have achieved, but by what you have overcome.” - Fannie Flagg

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Total population size estimation is an important problem in many social and biological sciences where data collection is a challenge. This task is performed using a special data collection procedure called capture-recapture. Capture-recapture was first used by Graunt (Krebs et al., 2014) to estimate the total population of England. Petersen (1896) was the first to apply this method on animal population. The basic set-up involves capturing a subset of animals from the population of interest, tagging and releasing them back into the population. This is followed by another round of capture, and hence, the term 'capture-recapture'. The two sets of capture are referred to as lists. This set-up is applicable to any total population size estimation problem where data is obtained from at least two different lists. Some of the well known instances of this problem include estimation of plague prevalence in England by Graunt (Hald, 2003), estimation of size fish (plaice) population (Petersen, 1896), estimation of number of pages on the world wide web (Fienberg et al., 1999) and estimation of number of victims of war (Ball et al., 2003; Lum and Ball, 2015; Manrique-Vallier et al., 2019). Over the years there have been several advancements and variations in this method to more accurately model the real data, which potentially increasing the complexities. *Continuation to be added...* In this thesis, we focus on the closed population set-up, i.e., where there are no additions or deletions in the population during the course of the data collection. Further, the capture occasions are discrete.

Total population size estimation is a missing data problem, where it is a challenge to appropriately infer about the size of the unobserved portion of the data from only the observed fraction. One has to use additional assumptions to ensure parameter identifiability. One has to be cautious when employing an assumption since, an invalid assumption will result in wrong inference. Using a strong assumption can give consistent estimates, however, if it is not valid for the data, then risk is higher. On the other hand using mild assumptions have lower risks of producing wrong estimates, but at the cost of higher uncertainty. Following this, one has to appropriately model the capture pattern in the data. Parametric models if correct can achieve good estimates with fast convergence rates. Similar to the the issue of identifiability, parametric assumptions can be too strong, especially when the data is complex. Use of flexible nonparametric models has gained popularity in recent literate, but it lacks some of the desirable properties of parametric models, like fast convergence rates and asymptotic normality. A detailed discussion of the existing approaches in the capture-recapture literature is available in chapter 2. Our work takes the nonparametric perspective, but uses advances in efficiency theory to characterize optimality and improve simple plug-in estimators (Bickel et al., 1993; van der Laan and Robins, 2003; Kennedy, 2016).

In chapter 2, under an identifying assumption that two lists are conditionally independent given measured covariates, we make several contributions. First, we derive the nonparametric efficiency bound for estimating the capture probability, which indicates the best possible performance of any estimator, and sheds light on the statistical limits of capture-recapture methods. Then we present a new estimator, that has a double robustness property new to capture-recapture, and is near-optimal in a non-asymptotic sense, under relatively mild nonparametric conditions. Next, we give a confidence interval construction method for total population size from generic capture probability estimators, and prove non-asymptotic near-validity. Finally, we apply them to estimate the number of killings and disappearances in Peru during its internal armed conflict between 1980 and 2000.

In chapter 3, we discuss a new R package `drpop` that implements the proposed method of chapter 2 on real data. `drpop` provides the user with the flexibility to choose the model for estimation of intermediate parameters and returns the estimated population size, confidence interval and some other related quantities. In this chapter, we illustrate the applications of `drpop` in different scenarios and we also present some performance summaries.

In chapter 4, we discuss a more relaxed set-up where we deviate from the conditional independence assumption. Under this-set-up, the target parameters are no longer point identified. Instead, we focus on estimating the identifiable range of these parameters. We used a sensitivity approach based on the conditional risk ratio between two lists in the presence of covariates and proposed confidence intervals using the formula of Imbens and Manski (2004). We have also evaluated the finite sample coverage guarantees of the general Imbens and Manski (2004) confidence interval. Finally, we present the performance results against the baseline plug-in approach in a simulated set-up and apply it to the Peru Internal Armed Conflict Data 1980-2000.

Chapter 2

Doubly robust capture-recapture methods for estimating total population size

2.1 Introduction

Capture-recapture is a study design for estimating population size when only a fraction of the population is observed. This setup arises frequently, for example in studying ecological abundance, disease prevalence, and casualties in armed conflicts. Capture-recapture has a long history, dating back to at least Graunt in the 1600s (Hald, 2003), who used it to estimate plague prevalence in England. Similarly, in 1802 Laplace estimated the total population of France (Goudie and Goudie, 2007), and Petersen (1896) the abundance of plaice fish. More recently, it has been used in diverse settings ranging from estimating the number of pages on the web (Fienberg et al., 1999) to the total number of victims in a war (Ball et al., 2003), among many others.

The simplest capture-recapture setup, credited to Petersen (1896); Lincoln (1930), consists of two independent lists with partial captures from the population of interest. There have been many generalizations over time. For our purposes, much of the previous work in capture-recapture can be viewed as falling within one of three streams. The first and oldest stream includes relatively simple data structures, e.g., involving no covariate information and relatively few lists (Petersen, 1896; Schnabel, 1938; Darroch, 1958). More recent advances in this stream include Burnham and Overton (1979) and Lee and Chao (1994). A second stream emerged to handle more intricate data structures, e.g., complex covariate information to help account for heterogeneity/dependence, largely using model-based approaches (Link, 2003; Carothers, 1973; Fienberg, 1972; Tilling and Sterne, 1999; Pollock, 2002; Huggins, 1989; Alho, 1990; Yip et al., 2001). However, the advantages of this second stream typically come at the expense of potentially restrictive parametric modeling assumptions, which when violated would induce bias. A third more recent stream addresses similar data structures as the second, but using more flexible nonparametric tools, e.g., local kernel or nonparametric Bayes or spline methods (Huggins and Hwang, 2007, 2011; Chen and Lloyd, 2000; Manrique-Vallier, 2016; Kurtz, 2018; Zwane and van der Heijden, 2005; Stoklosa and Huggins, 2012; Yee et al., 2015). However, the work in this third stream has so far relied on interpretable but typically suboptimal plug-in estimators, which can suffer from nonparametric smoothing bias and slow rates of convergence (van der Laan and Robins, 2003; van der Vaart, 2014; Robins et al., 2008). We refer to Kurtz (2018) for a more detailed review of this stream.

Our work takes the nonparametric perspective, but uses advances in efficiency theory to characterize optimality and improve simple plug-in estimators (Bickel et al., 1993; van der Laan and Robins, 2003; Kennedy, 2016). Under an identifying assumption that two lists are conditionally independent given measured covariate information (described in Section 2.2), we make several contributions.

- In Section 2.3 we derive the nonparametric efficiency bound for estimating the capture probability, which indicates the best possible performance of any estimator, and sheds light on the statistical limits of capture-recapture methods.
- In Section 2.4 we present a new doubly robust estimator, and study its finite-sample error, showing it is near-optimal in a non-asymptotic sense, under mild nonparametric conditions.
- In Section 2.5 we give a general method for constructing confidence intervals for population size from generic capture probability estimators, and prove non-asymptotic near-validity.
- In Section 2.6 we study our methods in simulations, and apply them to estimate the number of killings and disappearances attributable to different groups in Peru during its internal armed conflict between 1980 and 2000.

2.2 Preliminaries

2.2.1 Setup

Consider a finite population of n individuals, where the size n is unknown and to be estimated. We suppose there are K different lists of individuals from this population, yielding indicators $Y_{ik} \in \{0, 1\}$ of whether individual $i \in \{1, ..., n\}$ appeared on list $k \in \{1, ..., K\}$. We let $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK})^{\mathrm{T}}$ denote the vector indicating list membership (i.e., capture profile) information for individual i. For example, in the $K = 2$ case, a profile $\mathbf{Y}_i = (1, 0)^T$ would mean that individual i appears on list 1 but not list 2. We consider the case where covariates $\mathbf{X}_i \in \mathbb{R}^d$ are also available for each individual $i = 1, ..., n$. We assume an individual's chances of appearing on any given lists (and their covariates) do not depend on what happens with any other individuals, and also that the covariate and (conditional) list membership distributions are the same across individuals $i = 1, ..., n$. This implies that the random vectors $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i)$ are independent and identically distributed according to some distribution $\mathbb{P}$.

Remark 1. *The setup above is commonly referred to as "heterogeneous" (Huggins, 1989; Tilling and Sterne, 1999; Pollock, 2002) since list membership* $\mathbf{Y}$ *can vary with covariates* $\mathbf{X}$*. In other words, individuals with different covariates can have different chances of list membership.*

Remark 2. *In what follows we use the following standard notation. We let* $\mathbb{E}_{\mathbb{Q}}$ *denote an expectation under distribution* $\mathbb{Q}$*, and let* $\|f\|_{\mathbb{Q}}^2 = \int f(z)^2 \, d\mathbb{Q}(x)$ *denote the corresponding squared* $L_2(\mathbb{Q})$ *norm; we let* $\mathbb{Q}_N$ *denote the empirical measure under distribution* $\mathbb{Q}$*. Finally we let* $a \lesssim b$ *mean* $a \leq Cb$ *for some universal constant* C*.*

If every individual in the population appeared on at least one list (and could be uniquely identified), then the population size would of course be known without error; however in practice a possibly substantial fraction of individuals do not appear on any list. In other words, there are some individuals with $\mathbf{Y} = \mathbf{0}$ that we do not observe. This means that, although the distribution $\mathbb{P}$ governs the capture profiles, we cannot sample from $\mathbb{P}$ directly. Instead we only see the $N = \sum_{i=1}^{n} \mathbb{1}(\mathbf{Y}_i \neq \mathbf{0})$ individuals for whom $Y_{ik} = 1$ for some k. This is illustrated in Figure 2.1.

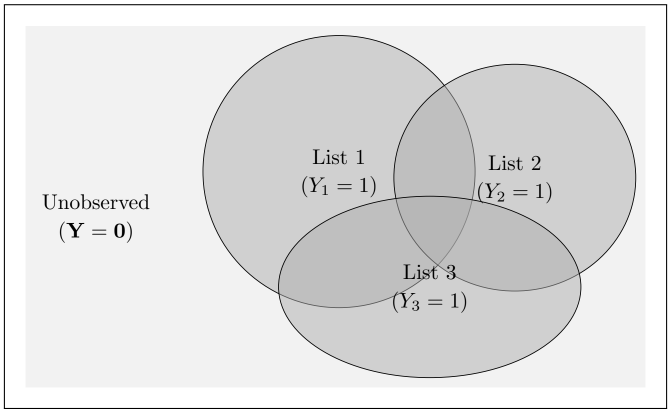

Figure 2.1: Schematic of data structure for $K = 3$ lists. Observed data (i.e., those with $\mathbf{Y} \neq \mathbf{0}$ in the union of the three lists) are represented with dark gray, while unobserved individuals with $\mathbf{Y} = \mathbf{0}$ are in light gray. Individuals appearing in all three lists have $\mathbf{Y} = (1, 1, 1)$.

Hence the capture-recapture design is an example of biased sampling (Vardi, 1985; Breslow et al., 2000; Qin, 2017). In particular, the observed data $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i), i = 1, ..., N$ are actually iid draws from a conditional distribution $\mathbb{Q}$ defined as

$$\begin{aligned}\mathbb{Q}(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}) &\equiv \mathbb{P}(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x} \mid \mathbf{Y} \neq \mathbf{0}) \\ &= \psi^{-1}\mathbb{P}(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x})\ \mathbb{1}(\mathbf{y} \neq \mathbf{0}) \end{aligned} \tag{2.1}$$

where ψ is the (marginal) capture probability defined as

$$\psi = \mathbb{P}(\mathbf{Y} \neq \mathbf{0}). \tag{2.2}$$

Remark 3. *Some authors use N to denote the total population size and n for the observed number of captures; in contrast, we use N for the observed number and n for the total size, following the convention of saving upper case for random variables. Note the observed number of captures N is random since it depends on the random selection indicators, while the total population size n is fixed; specifically $N \sim Bin(n, \psi)$. Nonetheless much of our analysis will be conditional on the observed sample size N.*

Recall our overall goal is to estimate the total population size $n = N + \sum_i \mathbb{1}(\mathbf{Y}_i = \mathbf{0})$. Since $N \sim \text{Bin}(n, \psi)$, the population size can be viewed as a fixed population parameter

$$n = \mathbb{E}(N)/\psi. \tag{2.3}$$

Intuitively, the lower the capture probability, the more the observed number N must be inflated to reflect the total population size. The quantity $\mathbb{E}(N)$ in (2.3) can of course be unbiasedly estimated with the observed number of captures N; therefore estimating population size essentially boils down to estimating the capture probability, which can then be used to inflate the observed N via the estimator

$$\widehat{n} = N/\widehat{\psi}. \tag{2.4}$$

Thus we turn towards the crucial question of how to efficiently estimate the capture probability (specifically, in the presence of high-dimensional and/or complex covariates $\mathbf{X}$), before coming back to inference about n in Section 2.5. In the next section we discuss identification of ψ from the distribution $\mathbb{Q}$ from which we sample; this will require extra assumptions, since if lists are irreparably dependent we will have no information about those who are unobserved.

Remark 4. *For the upcoming sections, we will only use the information of lists 1 and 2. Hence, for any captured individual, we only use the information $\mathbf{Z}_i = (Y_{i1}, Y_{i2}, \mathbf{X}_i)$. The captured individuals, that appear in neither list1 nor 2, have information vector $\mathbf{Z}_i = (0, 0, \mathbf{X}_i)$. More discussion follows in the next section.*

2.2.2 Identification

As mentioned in the previous section, without additional assumptions, the observed data distribution $\mathbb{Q}$ of list membership among those on at least one list is completely uninformative about the capture probability $\psi = \mathbb{P}(\mathbf{Y} \neq \mathbf{0})$. A variety of assumptions have been used to identify this and related quantities in previous literature; broadly, there must be some lack of dependence across lists in order to identify and estimate the odds and thus the overall population size. The popular Lincoln-Petersen estimator (Petersen, 1896; Lincoln, 1930) assumed independence between $K = 2$ lists and Darroch (1958) extended it to assume independence across $K > 2$ lists. Fienberg (1972) assumed a log-linear model for the expected number of observations with each capture profile, with one parameter necessarily set to zero (typically the highest-order interaction term across lists). You et al. (2021) have presented a generalizable framework adaptable to various identification assumptions including log-linear model assumption and independence between two lists conditional on the remaining list(s) for $K > 2$ case.

When one has access to not only list membership but also covariate information, conditional versions of these assumptions can be used (Sekar and Deming, 1949; Chao, 1987; Tilling and Sterne, 1999; Huggins and Hwang, 2007). Importantly, this allows heterogeneous capture probabilities that vary across units, and thus relaxes identifying assumptions; this is conceptually similar to how measured confounders are exploited in observational studies for causal inference (Hernan and Robins, 2019). Sekar and Deming (1949) studied the conditional case in the discrete and low-dimensional setup; the continuous case has been studied using parametric models by Pledger (2000); Pollock et al. (1990); Tilling and Sterne (1999); Huggins (1989); Chao (1987); Alho (1990). Burnham and Overton (1979); Huggins and Hwang (2007) used non-parametric jackknife estimator and bandwidth selection respectively. Analogous to the assumption of Tilling and Sterne (1999) for the two list case, our main identifying assumption for the K list case is that there is a known pair among the K lists which are conditionally independent. Without loss of generality, we order the lists so the first two are conditionally independent:

Assumption 1. $\mathbb{P}(Y_1 = 1 \mid \mathbf{X} = \mathbf{x}, Y_2 = 1) = \mathbb{P}(Y_1 = 1 \mid \mathbf{X} = \mathbf{x}, Y_2 = 0)$*, where* Y_k *denotes the capture indicator variable for list* k *for* $k = 1, \ldots, K$*.*

Assumption 1 says that the chance of appearing on list 1 is the same regardless of list 2 membership, among those with the same measured covariate values, i.e., that the list indicators Y_1 and Y_2 are conditionally independent given $\mathbf{X}$. This assumption can be viewed as an important relaxation of a more standard assumption of marginal independence, particularly when lists cover different parts of a population.

For example, consider a toy setup where there are two regions, equally populous. Suppose people who live in region A have a 90% chance of appearing on list 1 and a 10% chance of appearing on list 2, while people who live in region B have the reverse: a 10% chance of appearing on list 1 and a 90% chance on list 2. Thus list 1 tends to capture region A people, and list 2 tends to capture region B people. In this case, even if conditional independence holds (i.e., the chance of appearing on list 1, within each region, is the same regardless of whether you appear on list 2), the lists will *not* be marginally independent. Intuitively, the reasoning behind this is straightforward. Once we know you are on list 2, we have some additional information about your chances of appearing on list 1: you are more likely to live in region B, and so less likely to be on list 1. More specifically, the chances of appearing on either list are both 50% marginally, but the chance of appearing on list 1 given that you are on list 2 is only 18%.

The conditional independence in Assumption 1 has been used relatively extensively in the capture-recapture literature; we refer to Alho et al. (1993); Tilling and Sterne (1999); Tilling (2001); Brenner (1995); Pollock et al. (1990); Huggins (1989) for more details and discussions of when this assumption may hold and when it may fail. This assumption is violated for example when not all the covariates are measured. For example suppose the capture probabilities are influenced by location, age and gender of the individuals. If we measure only gender and age, we cannot assume conditional independence between lists 1 and 2. One can use tools like sensitivity analysis in this case. There is more discussion on this in section 2.7.

It is known (e.g., as in Tilling and Sterne, 1999) that under Assumption 1 the capture probability $\psi = \mathbb{P}(\mathbf{Y} \neq \mathbf{0})$ can be identified from the biased observed data distribution $\mathbb{Q}$. Specifically, let

$$\begin{aligned} q_1(\mathbf{x}) &= \mathbb{Q}(Y_1 = 1 \mid \mathbf{X} = \mathbf{x}) \\ q_2(\mathbf{x}) &= \mathbb{Q}(Y_2 = 1 \mid \mathbf{X} = \mathbf{x}) \\ q_{12}(\mathbf{x}) &= \mathbb{Q}(Y_1 = 1, Y_2 = 1 \mid \mathbf{X} = \mathbf{x}) \end{aligned}$$

denote the observational probability (under $\mathbb{Q}$) of appearing on list 1, 2, and both, respectively. These probabilities will be referred to as the q-probabilities throughout.

Remark 5. *Note that when there are only $K = 2$ lists, it must be that $q_1(\mathbf{x}) + q_2(\mathbf{x}) - q_{12}(\mathbf{x}) = 1$ since each observed unit must appear on list 1, list 2, or both, according to the sampling distribution $\mathbb{Q}$. In general when $K > 2$ it only holds that $0 \leq q_1(\mathbf{x}) + q_2(\mathbf{x}) - q_{12}(\mathbf{x}) \leq 1$, since some individuals may only appear on lists $j \geq 3$ other than 1 and 2.*

Remark 6. *Note that when there are more than two lists and without loss of generality the conditionally independent list pair is 1 and 2, the remaining lists aid the estimation by potentially increasing the number of observed individuals. We include the information of such individuals as shown in remark 4. Our method does not discard any row if they are captured neither in list 1 nor 2. This in turn leads to variance reduction as discussed in appendix A.1.1.*

For posterity we give the identification result for ψ in the following proposition (with a proof given in the Appendix).

Proposition 1. *Under Assumption 1 and the positivity condition* $\mathbb{Q}\{q_{12}(\mathbf{X}) > 0\} = 1$*, the conditional and marginal capture probabilities are identified from* $\mathbb{Q}$ *by*

$$\gamma(\mathbf{x}) \equiv \mathbb{P}(\mathbf{Y} \neq \mathbf{0} \mid \mathbf{X} = \mathbf{x}) = \frac{q_{12}(\mathbf{x})}{q_1(\mathbf{x})q_2(\mathbf{x})} \tag{2.5}$$

$$\psi \equiv \mathbb{P}(\mathbf{Y} \neq \mathbf{0}) = \left\{ \int \gamma(\mathbf{x})^{-1} \, d\mathbb{Q}(\mathbf{x}) \right\}^{-1}. \tag{2.6}$$

In the following sections we give three main contributions. First we derive the efficiency bound for estimating the capture probability ψ under a nonparametric model that puts no parametric restrictions on the 'nuisance' functions (q_1, q_2, q_{12}); second, we construct novel estimators that attain the efficiency bound under weak nonparametric conditions (e.g., allowing the use of flexible machine learning tools); and third, we give a general method for building corresponding confidence intervals for the total population size n, given any asymptotically linear estimate $\widehat{\psi}$ of the capture probability.

Remark 7. *All subsequent results apply to the statistical parameter* ψ*, which we define from here on as the harmonic mean on the right-hand-side of* (2.6)*. Under the identifying Assumption 1 (and positivity),* ψ *also represents the capture probability on the left-hand-side of* (2.6)*, but our statistical results do not require this link, and apply to the harmonic mean in* (2.6) *regardless.*

Remark 8. *When there are more than two lists, it is possible that multiple list pairs satisfy assumption 1. Assumption 1 is agnostic about the presence/absence of additional structure in the data. In the presence of extra structure, our approach, perhaps not the most efficient, is still valid. This presents multiple different opportunities to identify and estimate the capture probability, and so yields a semiparametric model with testable implications. We leave exploration of this setup for potential future work.*

2.3 Efficiency Bound

In this section, we derive the nonparametric efficiency bound for estimating the capture probability ψ using an iid sample from distribution $\mathbb{Q}$. This gives a crucial benchmark against which one can compare candidate estimators: once an estimator is shown to attain this bound, no further improvements can be made (at least asymptotically) without adding extra assumptions. To the best of our knowledge, the only previous efficiency bounds in the capture-recapture literature are for low-dimensional parametric models, where standard results from maximum likelihood theory apply (Fienberg, 1972; Gimenez et al., 2005).

In order to derive the efficiency bound, we use tools from semiparametric theory (Bickel et al., 1993). A fundamental goal here is to characterize influence functions, and the efficient influence function in particular. The efficient influence function of a parameter acts as the derivative term in a distributional Taylor expansion of the parameter, viewed as a map on distributions; thus it can represent the change in the parameter after perturbing the distribution it takes as input. Practically, the efficient influence function for a parameter has several critical implications. First, as mentioned above, it leads to a minimax efficiency bound (van der Vaart, 2002a) and thus provides a benchmark for efficient estimation in flexible nonparametric models. Further, it can be used to construct efficient estimators that attain the bound under weak assumptions, and sheds light on the regularity conditions necessary for said efficiency, as will be shown in Section 2.4. More details on nonparametric efficiency theory can be found in Bickel et al. (1993), van der Vaart (2002a), and van der Laan and Robins (2003), among others; reviews can be found in Tsiatis (2006) and Kennedy (2016), for example.

Our first result gives the form of the efficient influence function for the capture probability, in an unrestricted nonparametric model.

Lemma 2.0.1. *Let $g : \mathbb{R} \mapsto \mathbb{R}$ be any function differentiable at the true capture probability ψ defined in (2.6). Under a nonparametric model, the efficient influence function for the parameter $g(\psi)$ is given by $f_g(\psi)\phi(\mathbf{Z};\mathbb{Q})$ where $f_g(\psi) = -g'(\psi)\psi^2$ and*

$$\phi(\mathbf{Z};\mathbb{Q}) = \frac{1}{\gamma(\mathbf{X})}\left\{\frac{Y_1}{q_1(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right\} - \frac{1}{\psi}. \tag{2.7}$$

The proof is presented in the Appendix. Note that the first term in the efficient influence function is a product of the inverse conditional capture probability γ^{-1} with a term whose conditional expectation given $\mathbf{X}$ (under $\mathbb{Q}$) equals 1. The efficient influence function will be bounded for example if $q_{12}(\mathbf{x}) \geq \epsilon$ for some $\epsilon > 0$ and all $\mathbf{x}$ (note that $q_1(\mathbf{x}) \wedge q_2(\mathbf{x}) \geq q_{12}(\mathbf{x})$ so $q_{12}(\mathbf{x}) \geq \epsilon$ implies all the q-probabilities are bounded below by ϵ).

The variance of the efficient influence function acts as a nonparametric efficiency bound, in that no estimator can achieve a better mean squared error in a local minimax sense (van der Vaart, 2002b). The following theorem and corollary give the form of this bound and formalize the minimax result. All expectations and variances are under distribution $\mathbb{Q}$ unless noted otherwise.

Theorem 2.1. *Let $g : \mathbb{R} \mapsto \mathbb{R}$ be any function differentiable at ψ. The nonparametric efficiency bound for estimation of $g(\psi)$ is given by $var\{f_g(\psi)\phi(\mathbf{Z};\mathbb{Q})\} \equiv f_g(\psi)^2\sigma^2$, where $f_g(\psi)$ is defined in Lemma 2.0.1,*

$$\sigma^2 = \mathbb{E}\left(\frac{1}{\gamma(\mathbf{X})}\left[\left\{\frac{1-\gamma(\mathbf{X})}{\gamma(\mathbf{X})}\right\}\left\{\frac{1-q_{12}(\mathbf{X})}{q_{12}(\mathbf{X})}\right\} + \frac{q_0(\mathbf{X})}{q_{12}(\mathbf{X})}\right]\right) + var\left\{\frac{1}{\gamma(\mathbf{X})}\right\}$$

and $q_0(\mathbf{x}) = 1 - q_1(\mathbf{x}) - q_2(\mathbf{x}) + q_{12}(\mathbf{x})$ is the chance of appearing on neither list 1 nor 2.

The magnitude of the efficiency bound in Theorem 2.1 is driven by three main factors:

(i) the magnitude of the conditional capture probabilities,

(ii) the chance of appearing on both lists, and

(iii) the heterogeneity in the conditional capture probabilities.

Remark 9. *The efficiency bound for any function* $g(\cdot)$ *is always proportional to* σ^2*, with a scaling* $f_g(\psi)^2$ *depending on* g*; for example,* $f_g(\psi) = -1$ *when* $g(\psi) = 1/\psi$*, and* $f_g(\psi) = \psi/(1-\psi)$ *when* $g(\psi) = logit(\psi)$*. Therefore we focus our discussion on the quantity* σ^2*.*

The dependence on (i) in the bound in Theorem 2.1 occurs through the term $(1-\gamma)/\gamma^2$, i.e., the odds of capture divided by the capture probability. The dependence on (ii) occurs through the odds $(1-q_{12})/q_{12}$ as well as the probability ratio q_0/q_{12}. The dependence on the heterogeneity (iii) occurs through the $var(1/\gamma)$ term. Note that the probabilities γ and q_{12} in (i) and (ii) are related, but q_{12} can be small even when the capture probability γ is not, depending on the size of q_1 and q_2.

More specifically, all else equal, the variance bound increases with: (i) smaller capture probabilities γ, (ii) smaller chances of appearing on both lists q_{12}, and (iii) greater heterogeneity in the capture probabilities γ. Therefore capture probabilities can be estimated most efficiently when capture is likely, when there is substantial overlap across lists, and when capture probabilities are more homogeneous.

Remark 10. *For* $K = 2$*, the quantity* $q_0(\mathbf{x})$ *is exactly zero, but when* $K > 2$ *it can be positive.*

Remark 11. *When* $K = 2$ *and in the absence of covariates, the quantity* σ^2 *reduces to* $(\frac{1-\psi}{\psi^2})(\frac{1-q_{12}}{q_{12}})$*.*

In addition to informing what factors yield more or less efficient capture probability estimation, the variance in Theorem 2.1 also acts as a local minimax lower bound, as formalized in the following corollary.

Corollary 2.1.1. *For any estimator* $g(\widehat{\psi})$*, it follows that*

$$\inf_{\delta>0} \liminf_{N\to\infty} \sup_{TV(\overline{\mathbb{Q}},\mathbb{Q})<\delta} \frac{\mathbb{E}_{\overline{\mathbb{Q}}}\left[\{g(\widehat{\psi}) - g(\overline{\psi})\}^2\right]}{f_g(\psi)^2(\sigma^2/N)} \geq 1$$

where $TV(\overline{\mathbb{Q}},\mathbb{Q})$ *is the total variation between* $\overline{\mathbb{Q}}$ *and* $\mathbb{Q}$*,* $\psi = \psi(\mathbb{Q})$ *and* $\overline{\psi} = \psi(\overline{\mathbb{Q}})$ *are the capture probabilities under* $\mathbb{Q}$ *and* $\overline{\mathbb{Q}}$*, respectively, and* $f_g(\psi)$ *and* $\sigma^2 = \sigma^2(\mathbb{Q})$ *are defined as in Theorem 2.1.*

Corollary 2.1.1 shows that the worst-case mean squared error of *any* estimator of ψ, locally near the true $\mathbb{Q}$, cannot be smaller than the efficiency bound, asymptotically and after scaling by N. This local minimax result gives an important benchmark for efficient estimation of the inverse capture probability: no estimator can have mean squared error uniformly better than the variance of the efficient influence function divided by N, without adding extra assumptions and/or structure to the nonparametric model we consider.

Remark 12. *The local minimax result in Corollary 2.1.1 holds for any subconvex loss function* $\ell : \mathbb{R} \mapsto [0,\infty)$ *applied to* $\sqrt{N}\{g(\widehat{\psi}) - g(\overline{\psi})\}$*, not just squared error loss* $\ell(t) = t^2$*; the denominator lower bound in the general case is* $\mathbb{E}\{\ell(f_g\sigma Z)\}$ *where* $Z \sim \mathcal{N}(0,1)$ *is a standard normal random variable (van der Vaart, 2002b).*

Importantly, in the next section we construct estimators that can achieve the nonparametric efficiency bound under weak conditions that allow for flexible estimation of the q-probabilities, e.g., using machine learning tools.

2.4 Efficient Estimation

2.4.1 Setup

Recall we let $\mathbb{Q}_N$ denote the empirical measure under $\mathbb{Q}$, so that sample averages can be written with the short-hand $\mathbb{Q}_N(f) = \mathbb{Q}_N\{f(\mathbf{Z})\} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{Z}_i)$. The simplest estimator of the capture probability ψ is just a plug-in

$$\widehat{\psi}_{pi} = \left[\mathbb{Q}_N\left\{\frac{1}{\widehat{\gamma}(\mathbf{X})}\right\}\right]^{-1} \tag{2.8}$$

which replaces unknown quantities in the definition of ψ with estimates, i.e., by estimating the conditional capture probability $\widehat{\gamma}(\mathbf{X}) = \frac{\widehat{q}_{12}(\mathbf{X})}{\widehat{q}_1(\mathbf{X})\widehat{q}_2(\mathbf{X})}$ for every unit, and computing the harmonic mean of the values across the sample. This estimator has been used relatively extensively in previous work (Huggins and Hwang, 2007; Darroch, 1958; Fienberg, 1972; Tilling and Sterne, 1999). When the q-probabilities are estimated with correctly specified parametric models, the plug-in estimator $\widehat{\psi}_{pi}$ will be $\sqrt{n}$-consistent and asymptotically normal under standard regularity conditions. However, when the covariates contain any continuous components and/or are high-dimensional, it is usually very unlikely an analyst would have enough a priori knowledge to be able to correctly specify a low-dimensional parametric model, let alone three (one for each q-probability nuisance function).

This difficulty of correct model specification suggests trying to flexibly estimate the q-probabilities, e.g., using logistic regression with model selection, or lasso, or nonparametric tools like random forests, neural nets, RKHS regression, etc. Unfortunately, when the plug-in estimator $\widehat{\psi}_{pi}$ is constructed from these kinds of data-adaptive methods, it in general loses the nice properties it has in the parametric setup. Specifically, without special tuning of particular methods, it in general would suffer from slower than $\sqrt{n}$-convergence rates, and have an unknown limiting distribution, making it not only inefficient but also leaving no tractable way to do inference. This deficiency of plug-in estimators is by now relatively well-known in functional estimation problems (van der Laan and Robins, 2003; Chernozhukov et al., 2018; Wu et al., 2019); however we have not seen it highlighted in the capture-recapture setting. (We show these issues via simulations in Section 2.6.1.)

Luckily, the plug-in can be improved upon using tools from semiparametric efficiency theory (Bickel et al., 1993; van der Vaart, 2002a; van der Laan and Robins, 2003; Tsiatis, 2006; Kennedy, 2016). In what follows, we will present and study a novel doubly robust estimator, which can attain the efficiency bound from the previous section even when built from flexible data-adaptive regression tools.

2.4.2 Doubly Robust Estimator

As mentioned above, the plug-in estimator (2.8) has some important deficiencies in semi- and non-parametric settings. The plug-in estimator can be debiased by adding an estimate of the mean of the efficient influence function (Bickel et al., 1993; van der Vaart, 2002a; van der Laan and Robins, 2003; Tsiatis, 2006; Kennedy, 2016). This leads to our proposed doubly robust estimator

$$\widehat{\psi}_{dr} = \mathbb{Q}_N \left[\frac{1}{\widehat{\gamma}(\mathbf{X})} \left\{ \frac{Y_1}{\widehat{q}_1(\mathbf{X})} + \frac{Y_2}{\widehat{q}_2(\mathbf{X})} - \frac{Y_1 Y_2}{\widehat{q}_{12}(\mathbf{X})} \right\} \right]^{-1} \tag{2.9}$$

where $\widehat{q}_j$ are estimates of the q-probabilities (e.g., via regression predictions).

Remark 13. *In order to avoid potentially restrictive empirical process conditions, we estimate $\mathbb{Q}_N$ and $\widehat{q}_j$ from separate independent samples. Specifically, we estimate the q-probability nuisance functions by fitting regressions in a training sample, independent of a test sample $\mathbb{Q}_N$. With iid data, one can always obtain such samples by splitting at random in half, or folds. This yields a loss in efficiency, but that can be fixed by swapping the samples/folds, computing the estimate on each, and averaging. This is referred to as cross-fitting, and has been used for example by Bickel and Ritov (1988); Robins et al. (2008); Zheng and van der Laan (2010); Chernozhukov et al. (2017). Here we analyze a single split procedure, merely to simplify notation; extending to averages across independent splits is straightforward.*

In the following sub-section, we derive finite-sample error bounds and distributional approximations for our doubly robust estimator, which are valid for any sample size.

Non-asymptotic Error Bounds and Approximate Normality

In this section, we provide our three main theoretical results regarding error bounds for our proposed method. In particular we show that our estimator is nearly efficient, doubly robust, and approximately normal. Importantly, we show all these properties hold in finite samples, without resorting to asymptotics.

In the previous section, we derived the efficient influence function, which is the crucial component of the local minimax lower bound we gave in Corollary 2.1.1. This corollary shows the minimax optimal estimator has mean squared error that scales like the variance of the efficient influence function divided by N, so that an optimal estimator would be one that behaves like a sample average of the efficient influence function. Our first result shows that our proposed estimator *does* in fact behave like an average of the efficient influence function, depending on the size of nuisance error. In what follows, we use ϕ and $\widehat{\phi}$ to denote the efficient influence function $\phi(\mathbf{Z}; \mathbb{Q})$ and its estimate $\phi(\mathbf{Z}; \widehat{\mathbb{Q}})$ respectively.

Theorem 2.2. *For any sample size N and error tolerance $\delta > 0$, we have*

$$|(\widehat{\psi}_{dr}^{-1} - \psi^{-1}) - \mathbb{Q}_N \phi| \leq \delta$$

with probability at least

$$1 - \left(\frac{1}{\delta^2}\right) \mathbb{E}\left(\widehat{R}_2^2 + \frac{\|\widehat{\phi} - \phi\|^2}{N}\right)$$

where $\widehat{R}_2$ is a second-order error term given by

$$\begin{aligned}\widehat{R}_2 &= \int \frac{1}{\widehat{q}_{12}} \left\{ \left(q_1 - \widehat{q}_1\right)\left(\widehat{q}_2 - q_2\right) + \left(q_{12} - \widehat{q}_{12}\right)\left(\frac{1}{\gamma} - \frac{1}{\widehat{\gamma}}\right) \right\} \, d\mathbb{Q} \\ &\leq \left(\frac{1}{\epsilon}\right) \|\widehat{q}_1 - q_1\| \|\widehat{q}_2 - q_2\| + \left(\frac{1}{\epsilon^3}\right) \|\widehat{q}_{12} - q_{12}\| \|\widehat{\gamma} - \gamma\|\end{aligned}$$

with the latter bound on $\widehat{R}_2$ holding as long as $(q_{12} \wedge \widehat{q}_{12}) \geq \epsilon$.

Theorem 2.2 shows that our proposed estimator is within δ of a sample average of the efficient influence function, centered at the true (inverse) capture probability, with high probability, at every sample size. For a given observed number of captures N and error δ, this probability depends on two factors: (i) a second-order error term $\widehat{R}_2$, which is driven by the error in estimating the nuisance q-probabilities; and (ii) the L_2 error in estimating the efficient influence function itself, divided by N, which also depends on estimation error of the q-probabilities, but in a weaker way due to the division by N. When $\widehat{R}_2$ goes to 0 as N increases, the probability above goes to 1 for any fixed δ. Hence, Theorem 2.2 implies usual asymptotic convergence in probability, but in addition it gives an error bound that is valid for any finite N.

For example, if the q-probabilities are estimated with errors upper bounded by $cN^{-1/4}$, then with at least 95% probability, our proposed estimator will be within $4c^2\sqrt{5/N}$ of the average efficient influence function. More generally, if $\mathbb{E}|\widehat{R}_2| \lesssim 1/\sqrt{N}$, then our proposed estimator will be within $1/\sqrt{N}$ (up to constants) of this efficient average, with high probability. For example, if q_1, q_2, q_{12} and γ belong to Holder classes $\mathcal{H}(\beta_1)$, $\mathcal{H}(\beta_2)$, $\mathcal{H}(\beta_{12})$ and $\mathcal{H}(\beta_0)$ respectively, where $\mathcal{H}(s)$ is a Holder class with smoothness index s (Györfi et al., 2006; Tsybakov, 2008), then a sufficient condition for this kind of result, if the q-probabilities are estimated at minimax optimal rates, would be that the minimum smoothness is at least half the dimension of the covariates, i.e., $min\{\beta_0, \beta_{12}, \beta_1, \beta_2\} \geq d/2$. Similarly, if the q-probabilities were s-sparse, then a sufficient condition would be that $s \lesssim \sqrt{n}$ with lasso-style methods (Farrell, 2015). However, such $N^{-1/4}$ nuisance errors are only sufficient conditions for $1/\sqrt{N}$ capture probability errors; one only needs the remainder error $\widehat{R}_2$ to be small enough, which could also be achieved if some combinations of q-probabilities are estimated well, even if others are not. We give more detail on this phenomenon in our next result.

Namely, beyond being close to an optimally efficient estimator, in the next result we show that our proposed estimator enjoys a finite-sample multiple robustness phenomenon, never before shown in capture-recapture problems. This phenomenon indicates that the overall estimation error can be small as long as some, but not all, nuisance probabilities are estimated with small error.

Corollary 2.2.1. *Suppose $(q_{12} \wedge \widehat{q}_{12}) \geq \epsilon$. Assume one of the following holds:*

1. $\|\widehat{q}_1 - q_1\| \vee \|\widehat{q}_{12} - q_{12}\| \leq \xi_N$, *or*
2. $\|\widehat{q}_1 - q_1\| \vee \|\widehat{\gamma} - \gamma\| \leq \xi_N$, *or*
3. $\|\widehat{q}_2 - q_2\| \vee \|\widehat{q}_{12} - q_{12}\| \leq \xi_N$, *or*
4. $\|\widehat{q}_2 - q_2\| \vee \|\widehat{\gamma} - \gamma\| \leq \xi_N$.

Then $|(\widehat{\psi}_{dr}^{-1} - \psi^{-1}) - \mathbb{Q}_N \phi| \leq \delta$ *with probability at least*

$$1 - \left(\frac{C}{\delta^2}\right)\left(\xi_N^2 + \frac{1}{N}\right)$$

where C *is a constant independent of sample size* N*.*

As a consequence of Corollary 2.2.1, the proposed estimator is doubly robust, i.e., if either estimator of q_1 or q_2 has small error, and either estimator of q_{12} and γ has small error, then the overall error of our proposed estimator (given by $\widehat{R}_2$) will be just as small, up to constants, even if the other estimators have large errors or are misspecified. (We note that, although this kind of robustness is sometimes called multiple robustness (Vansteelandt et al., 2008), we use the term doubly robust since the error structure is still second-order, i.e., involving products of errors, albeit with more terms). This property is very useful when one of the lists is difficult to estimate, for example, due to high-dimensional covariates, or q-probabilities that are complex functions of continuous covariates.

Remark 14. *Note that when there are only* $K = 2$ *lists, then we have the relation* $q_1(\mathbf{x}) + q_2(\mathbf{x}) - q_{12}(\mathbf{x}) = 1$ *for all* $\mathbf{x}$*. Hence, small bias for any two estimators of* q_{12}*,* q_1 *and* q_2 *automatically implies small bias for the third. One might expect then that double robustness does not arise in the* $K = 2$ *list setting; however this is not quite true. To see why, note that it could be possible to estimate* γ *with small error, for example, even when some of the q-probability estimators are misspecified. These and other issues related to estimation of the conditional capture probability* γ *will be important to explore in future work.*

When the remainder error $\widehat{R}_2$ is sufficiently small, Theorem 2.2 and Corollary 2.2.1 tell us we can approximate $\widehat{\psi}_{dr}^{-1}$ with a sample average of the efficient influence function. For the purposes of inference, this suggests a confidence interval of the form

$$\widehat{\mathrm{CI}} = [\widehat{\psi}_{dr}^{-1} \pm z_{1-\alpha/2}\widehat{\sigma}/\sqrt{N}], \tag{2.10}$$

where $\widehat{\sigma}$ is a variance term defined in Theorem 2.3. In the next Berry-Esseen-type result, we exploit this closeness with a sample average and further show that our proposed estimator, properly scaled, is approximately Gaussian. This will show that the above confidence interval gives nearly-valid finite-sample coverage guarantees.

Theorem 2.3. *Let* $\widehat{\sigma}^2 = \widehat{var}(\widehat{\phi})$ *be the unbiased empirical variance of the estimated efficient influence function. Then* $\widehat{\psi}_{dr}^{-1} - \psi^{-1}$ *follows an approximately Gaussian distribution, with the difference in cumulative distribution functions uniformly bounded above by*

$$\left| \mathbb{P}\left(\frac{\widehat{\psi}_{dr}^{-1} - \psi^{-1}}{\widehat{\sigma}/\sqrt{N}} \leq t \right) - \Phi(t) \right| \leq \frac{C}{\sqrt{N}} \mathbb{E}\left(\frac{\rho}{\widetilde{\sigma}^3} \right) + \frac{1}{\sqrt{2\pi}} \left\{ \sqrt{N} \mathbb{E}\left(\frac{|\widehat{R}_2|}{\widetilde{\sigma}} \right) + |t| \mathbb{E}\left(\left| \frac{\widehat{\sigma}}{\widetilde{\sigma}} - 1 \right| \right) \right\} \tag{2.11}$$

where $\widetilde{\sigma} = var(\widehat{\phi} | \mathbf{Z}^n)$, $\rho = \mathbb{E}\{ |\widehat{\phi} - \mathbb{Q}\widehat{\phi}|^3 | \mathbf{Z}^n \}$ *and* $C < 1/2$ *is the Berry-Esseen constant.*

The above result shows that the estimation error scaled by $\widehat{\sigma}/\sqrt{N}$ is approximately standard normal. The first term on the right hand side of (2.11) is the usual Berry-Esseen bound. The second term captures the effect of the nuisance estimation error $\widehat{R}_2$. The third term is the estimation error in the variance. Since $\mathbb{E}|\widehat{\sigma} - \widetilde{\sigma}|$ is bounded above by $cN^{-1/2}$ (proof in the appendix), the overall error in the Gaussian approximation is driven by the second term, involving nuisance error $\widehat{R}_2$. This will be the main driver of whether the interval has approximately correct coverage. We note that the above theorem implies convergence in distribution whenever $\mathbb{E}|\widehat{R}_2| = o(1/\sqrt{N})$ (which can hold for a wide variety of flexible nonparametric estimators of the q-probabilities, as discussed after Theorem 2.2), but in addition gives a more precise error bound that holds for any finite sample size.

Note that Theorem 2.3 immediately implies that the error in coverage

$$\left| \mathbb{P}\left(\widehat{\text{CI}} \ni \psi^{-1} \right) - (1 - \alpha) \right|$$

for the proposed confidence interval defined in (2.10) is no more than twice the error bound on the right-hand-side of (2.11), with $t = z_{\alpha/2}$. Further, a Berry-Esseen-style bound similar to that of Theorem 2.3 (along with subsequent coverage guarantees and corollaries) can be obtained for any function $g(\cdot)$ of $\widehat{\psi}_{dr}^{-1}$ satisfying the conditions from Friedrich (1989). This implies the same kind of coverage guarantees for ψ, for example, using the confidence interval

$$\widehat{\psi}_{dr} \pm z_{1-\alpha/2} \widehat{\sigma} \widehat{\psi}_{dr}^2 / \sqrt{N}$$

which can be motivated via the delta method. The error in the coverage of this estimated interval is twice the bound in Theorem 2.3, modulo some extra dependence on g.

Importantly, the unbiased empirical variance $\widehat{\sigma}^2$ is a consistent estimator of the efficiency bound $\sigma^2 = var(\phi)$ in the sense that $\mathbb{E}|\sigma - \widehat{\sigma}| \lesssim \mathbb{E}\|\widehat{\phi} - \phi\| + N^{-1/2}$. This shows the crucial result that our estimator is approximately minimax optimal in the sense of Corollary 2.1.1, if the nuisance error is small enough.

A natural consequence of the above theorem is the following corollary, which presents a simple bound on the error of the normal approximation, under some natural conditions on the nuisance error $\widehat{R}_2$ and variance.

Corollary 2.3.1. *Assume* $\widetilde{\sigma} \gtrsim 1$, $\mathbb{E}|\widehat{R}_2| \lesssim N^{-2\beta}$ *and* $\alpha > \delta$ *for some* $\delta > 0$. *Then the coverage error for the proposed* $(1-\alpha)$ *confidence interval defined in* (2.10) *is upper bounded by*

$$\left|\mathbb{P}\left(\widehat{CI} \ni \psi^{-1}\right) - (1-\alpha)\right| \lesssim N^{(1-4\beta)/2} + \frac{1}{\sqrt{N}}.$$

Therefore if $\beta > 1/4$ *there exists some sample size* N_ϵ *at which the coverage error is never more than* ϵ, *for any* $N > N_\epsilon$.

Since this corollary is a special case of Theorem 2.3, mainly aimed at presenting the result in a simple form, we refer to the above discussion for more details. However we note that the condition that $\mathbb{E}|\widehat{R}_2| \lesssim N^{-2\beta}$ would hold for example if the q-probabilities were estimated optimally when contained in Holder classes with smoothness index s, where $\beta = \frac{s}{2s+d}$ (or under some conditions on sparsity, as discussed after Theorem 2.2). Then $\beta > 1/4$ would mean $s > d/2$, aligning with our earlier results.

In this section, we have given finite-sample error bounds and distributional approximations for our proposed estimator, which are valid for any sample size, allowing accurate estimation and approximately valid confidence guarantees, even in complex nonparametric models where the q-probabilities are estimated with flexible machine learning tools. In the next section, we consider a slightly modified version of the estimator which could further improve finite-sample properties.

2.4.3 Targeted Maximum Likelihood Estimator

We have seen that our proposed doubly robust estimator (2.9) is close in a finite-sample sense to an optimal sample average, and possesses crucial double robustness properties. However it is possible this estimator may not respect the bounds on the parameter space; for example $\widehat{\psi}_{dr}$ may fall outside $[0,1]$ if some of the estimates of the q-probabilities are small. A simple fix is to truncate the estimator $\widehat{\psi}_{dr}$ to always lie in $[0,1]$. Here for completeness we discuss an alternative approach using targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin, 2006; van der Laan and Rose, 2011), which is an iterative procedure that fluctuates nuisance estimates so that a plug-in estimator built from them also approximately solves an efficient influence function estimating equation. TMLE thus leads to estimators that are asymptotically equivalent to one-step bias-corrected estimators, but which could bring some finite-sample advantages.

In Appendix A.2, we present an algorithm (Algorithm 1) detailing the computation of a TMLE for ψ. At a high level, the procedure involves bias correction via iterative updating of initial nuisance estimates, based on quantities called clever covariates in the TMLE literature. Interestingly, in addition to being somewhat more computationally intensive, TMLE estimators are not sample averages like our main proposed estimator from the previous subsection; this makes it less clear how to derive finite-sample error bounds. Since the estimates $\widehat{q}_j^*(\mathbf{x})$ obtained after convergence satisfy $\mathbb{Q}_N\{\phi(\mathbf{Z};\widehat{\mathbb{Q}}^*)\} \approx 0$, the asymptotic behavior matches the doubly robust estimator in (2.9), but for describing finite-sample behavior we resort to simulations, detailed in Section 2.6.1.

2.5 Inference for Population Size

In the previous section we gave doubly robust estimators for the capture probability and studied finite-sample properties. In this section we give a crucial result that shows how to obtain an approximate confidence interval for the *population size*, given a generic initial estimator of the (inverse) capture probability. Importantly, our results only require this initial estimator to be weakly approximated by a sample average, and otherwise are completely agnostic to how the capture probability is estimated. This appears in stark contrast to most of the literature on this topic, where the inferential procedures are very closely tied to specific model assumptions and estimator constructions.

This main inferential result is given in the following theorem.

Theorem 2.4. *Suppose we are given an initial estimator* $\widehat{\psi}$ *that satisfies*

$$\widehat{\psi}^{-1} - \psi^{-1} = \mathbb{Q}_N\left(\widehat{\varphi}\right) - \int \widehat{\varphi}(\mathbf{z})d\mathbb{Q}(\mathbf{z}) + \widehat{R}_2$$

for φ *a generic influence function with mean zero and* $\widehat{R}_2$ *an error term. Let* $\widehat{\tau}^2 = \widehat{\psi}\widehat{\varsigma}^2 + \frac{1-\widehat{\psi}}{\widehat{\psi}}$ *and* $\widetilde{\tau}^2 = \psi\widetilde{\varsigma}^2 + \frac{1-\psi}{\psi}(\psi\widehat{R}_2+1)^2$*, where* $\widehat{\varsigma}^2 = \widehat{var}(\widehat{\varphi})$ *is the unbiased empirical variance of the estimated influence function and* $\widetilde{\varsigma} = var(\widehat{\varphi} \mid \mathbf{Z}^n)$ *the true conditional variance. Then the* $(1-\alpha)$ *confidence interval given by*

$$\widehat{CI}_n = \left[\widehat{n} \pm z_{\alpha/2}\widehat{\tau}\sqrt{\widehat{n}}\right] \tag{2.12}$$

has coverage error upper bounded as

$$\left|\mathbb{P}\left(\widehat{CI}_n \ni n\right) - (1-\alpha)\right| \leq \frac{2C}{\sqrt{n}}\mathbb{E}\left(\frac{\rho}{\widetilde{\tau}^3}\right) + \sqrt{\frac{2}{\pi}}\left\{\sqrt{n}\psi\mathbb{E}\left(\frac{|\widehat{R}_2|}{\widetilde{\tau}}\right) + |z_{\alpha/2}|\mathbb{E}\left(\left|\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - 1\right|\right)\right\} \tag{2.13}$$

where C *is the Berry-Esseen constant and*
$\rho = \mathbb{E}\left[\left|\mathbb{1}(\mathbf{Y} \neq \mathbf{0})\left\{\widehat{\varphi} - \mathbb{Q}\widehat{\varphi}\right\} + \left\{\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) - \psi\right\}\widehat{R}_2 + \psi^{-1}\left\{\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) - \psi\right\}\right|^3 \middle| \mathbf{Z}^n\right].$

Theorem 2.4 gives a non-asymptotic upper bound on how much the coverage $\mathbb{P}(\widehat{\mathrm{CI}}_n \ni n)$ of our proposed interval

$$\frac{N}{\widehat{\psi}} \pm z_{\alpha/2}\sqrt{\left(\widehat{\psi}\widehat{\varsigma}^2 + \frac{1-\widehat{\psi}}{\widehat{\psi}}\right)\frac{N}{\widehat{\psi}}}$$

can deviate from its nominal $(1-\alpha)$ level. Before describing the coverage guarantee, we first describe the proposed interval. The length of this interval is driven by three factors: (i) the estimated odds of not being captured $(1-\widehat{\psi})/\widehat{\psi}$, (ii) the variance of the inverse capture probability estimator $\widehat{\varsigma}^2$, and (iii) the sample size N. As one would expect, higher odds of capture yield more precise inference about population size, all else equal, as does more efficient estimation of $\widehat{\psi}$. Specifically, the length of the interval shrinks to zero when the capture probability is very large, regardless of the sample size N. Also note that even if ψ were known, one would still have an interval of the form

$$\frac{N}{\psi} \pm z_{\alpha/2}\sqrt{\frac{1-\psi}{\psi}}\sqrt{\frac{N}{\psi}}$$

based on the fact that $\widehat{n} = N/\sqrt{n}$ is approximately normal. Although sample size N appears in the numerator of the interval width (contrary to standard intervals), it only appears through its square root, showing that in an asymptotic regime where $N \to \infty$, the width still grows at a slower rate than the sample size. Intuitively, this interval takes $\widehat{n} = N/\widehat{\psi}$ and multiplies by $1 \pm z_{\alpha/2}/\sqrt{\widehat{n}}$, which does tend to zero as sample size N grows.

Remark 15. *For $K = 2$ lists and in the absence of covariates, the confidence interval reduces to $\widehat{n} \pm z_{\alpha/2}\sqrt{\frac{\widehat{n}(1-\widehat{\psi})}{\widehat{\psi}\;\widehat{q}_{12}}}$, which approximately resembles the Wald-type confidence interval for the Lincoln-Petersen estimator (Evans et al., 1996).*

Now we describe the coverage guarantee of Theorem 2.4. Importantly, the bound on the coverage error depends on a number of factors, as shown above appearing the sum of the three terms in (2.13). Under typical boundedness assumptions, the first and third terms would be of smaller order, and the second term would dominate. This second term is driven by the size of $\widehat{R}_2$ in terms of its mean absolute value, i.e., how well the initial estimator $\widehat{\psi}$ is approximated by a sample average. If $\widehat{R}_2$ is not substantially smaller than $1/\sqrt{n}$, then the confidence interval would not be guaranteed to cover the true population size n at its nominal level. This points to the importance of efficient estimation of ψ; for example, as shown in the previous section, our proposed estimator $\widehat{\psi}_{dr}$ can be approximated by a sample average up to smaller than $1/\sqrt{n}$ error, even in a nonparametric model when q-probabilities are estimated flexibly.

Remark 16. *A unique feature of Theorem 2.4 is that it is valid for* any *estimator approximated by a sample average, regardless of what underlying identification or estimation assumptions were used in its construction. This means if another analyst did not believe the independent lists condition in Assumption 1, and instead constructed an estimate of the capture probability under a different identifying assumption, they could also use the above theorem to construct a confidence interval and assess its finite-sample coverage.*

A natural consequence of Theorem 2.4 is the following corollary, which parallels Corollary 2.3.1 in giving a simple bound on normal approximation error, under natural conditions.

Corollary 2.4.1. *Assume* $\widetilde{\tau} \gtrsim 1$, $\mathbb{E}|\widehat{R}_2| \lesssim N^{-2\beta}$ *and* $\alpha > \delta$ *for some* $\delta > 0$. *Then the coverage error for the proposed* $(1-\alpha)$ *confidence interval defined in* (2.10) *is upper bounded by*

$$\left|\mathbb{P}\left(\widehat{CI}_n \ni n\right) - (1-\alpha)\right| \lesssim n^{(1-4\beta)/2} + \frac{1}{\sqrt{n}}.$$

Therefore if $\beta > 1/4$ *there exists some population size* n_ϵ *at which the coverage error is never more than* ϵ*, for any* $n > n_\epsilon$.

Since the result in Corollary 2.4.1 is similar to that of Corollary 2.3.1, we refer there for related discussion. The main point is that, as long as our initial estimator is well-approximated by a sample average, no matter how it was constructed or what assumptions it relies on, our proposed confidence interval (2.12) will be approximately valid.

2.6 Simulation & Application

So far we have proposed doubly robust estimators for the capture probability, and a general approach for constructing confidence intervals for the total population size, all with non-asymptotic error guarantees. In this section we study the performance of our methods in simulated data, and apply them to estimate the total number of killings in the internal armed conflict in Peru during 1980-2000. The code used to generate the results is available on github at `mqnjqrid/capture_recapture`.

2.6.1 Simulation

Here we use simulations similar to Tilling and Sterne (1999), taking $n = 5000$ samples from

$$\begin{aligned} &X \sim \mathit{Uniform}(2,3) \\ &\mathbb{P}(Y_1 = 1 \mid X = x) = \operatorname{expit}(a + 0.4x) \\ &\mathbb{P}(Y_2 = 1 \mid X = x) = \operatorname{expit}(a + 0.3x). \end{aligned}$$

where a takes values $\{-2.513, -0.66\}$ to ensure that the capture probability ψ takes values $\{0.3, 0.8\}$, respectively. This gives sample sizes N approximately equal to $\{1500, 4000\}$. Recall that under $\mathbb{P}$, list membership Y_1 and Y_2 are conditionally independent, so the conditional capture probability is

$$\gamma(x) = 1 - \{1 - \operatorname{expit}(a + 0.4x)\}\{1 - \operatorname{expit}(a + 0.3x)\}$$

and q-probabilities are equal to $q_j(x) = \mathbb{P}(Y_j = 1 \mid X = x)/\gamma(x)$.

We construct estimates of the q-probabilities via $\widehat{q}_j(x) = \text{expit}[\text{logit}\{q_j(x)\} + \epsilon_j]$, where we simulate the errors in estimation by $\epsilon \sim \mathcal{N}(n^{-\alpha}, n^{-2\alpha})$. This allows us to carefully control the error of the q-probability estimators; since the root mean squared error scales like $n^{-\alpha}$, this can be viewed as the rate of convergence. We run 500 simulations for each $\alpha \in \{0.1, 0.2, 0.25, 0.3, 0.4, 0.5\}$. Note α values 0.5 and 0.25 correspond to the parametric ($n^{-1/2}$) and nonparametric ($n^{-1/4}$) rates, respectively. Figure 2.2 shows the estimated bias and the root mean square error (RMSE) of $\widehat{\psi}$, along with the coverage proportion for the confidence interval of the total population size.

Remark 17. *For plug-in estimators, there is no well-defined variance formula (this is a main motivation for our doubly robust construction). Therefore to construct confidence intervals with the plug-in estimator, we used the estimated variance of the doubly robust estimator.*

Overall, the simulations illustrate the phenomena expected from our theoretical results: when the q-probabilities are estimated with low error (i.e., α large), all the methods do well, whereas when the q-probabilities are difficult to estimate (i.e., α smaller) the proposed methods do substantially better in terms of bias, error, and coverage. For example, when the true capture probability is 50%, the simple plug-in estimator gives substantial bias as soon as $\alpha < 0.4$ (i.e., when the q-probabilities are estimated at slower than $n^{-2/5}$ rates). However, the bias of the proposed doubly robust estimator is relatively unaffected until $\alpha < 0.2$, with the TMLE somewhere in between. The story is similar for the RMSE, which is largely driven by the bias in this problem. The coverage is approximately at the nominal 95% level as soon as $\alpha > 0.2$ (i.e., when the q-probabilities are estimated at faster than $n^{-1/5}$ rates), whereas the plug-in estimator substantially under-covers (e.g., nearly zero at $\alpha = 0.2$) until $\alpha \geq 0.4$. Using simulated data with capture probability 0.5, one will get results similar those of $\psi = 0.3$. We note that when population size or capture probability is small (e.g., capture probability substantially less than 50%), estimation becomes more challenging and the story is less clear about which method does better. For reference, results for population sizes varying from $n = 200$ to $n = 1000$ (in the $\alpha = 0.25$ case) are given in the Appendix in Figure A.1.

2.6.2 Data Analysis

We apply our proposed methods to estimate the number of killings and disappearances attributable to different groups in Peru during its internal armed conflict between 1980 and 2000. We use data collected by the Truth and Reconciliation commission of Peru (Ball et al., 2003), as well as detailed geographic information, following Rendon (2019a).

There is an ongoing debate regarding the total number of killings and disappearances in the conflict, as well as about which groups are most responsible, e.g., the PCP-Shining Path versus the State or other groups. Ball et al. (2003) estimated approximately 69,000 total killings and disappearences, finding the Shining Path responsible for the majority. In contrast, Rendon (2019a) estimated approximately 48,000 killings and disappearances, with the State responsible for the majority, though many geographic strata were excluded. Most recently, Manrique-Vallier et al. (2019) included a newly available list and estimated approximately 58,000-65,000 killings and disappearances, depending on choices of priors, with the Shining Path responsible for the majority. Before describing our specific approach, we first describe the data and give some summary statistics. As explained in Ball et al. (2003), the data come from a few main sources: the Truth and Reconciliation Commission (CVR), the Public Defender Office (DP), and 4–5 other human rights groups and NGOs (ODH). We use the CVR as our first list and construct the second list by combining the remaining lists, i.e., DP and ODH since they have similar demographics. The data contains identifiers of people who have been killed or disappeared, as well as which of the source lists they appeared on, and covariates including age, gender, and geographic location of the killing or disappearance (measured via 58 geographic strata as in Ball et al. (2003), as well as bivariate latitude/longitude as in Rendon (2019a)). To avoid missing completely at random assumptions, we also included missingness indicators for victims with missing age (28% missing), gender (¡1% missing), or location (11% missing). The lists of all the covariates is available in the appendix A.4. The total number of killings and disappearances across all lists was 24,692. Importantly, the lists capture different demographics, which points to the necessity of relaxing classical marginal independence via the *conditional* independence in Assumption 1. For example, the CVR list mostly includes victims who were killed, while the DP and ODH lists mostly include victims who disappeared, as shown in Figure 2.3. Similarly, geographic diversity varies across lists, as shown in Figure 2.4. For example, almost 60% of Shining Path victims in the DP and ODH lists come from two smaller districts (Chungui and Luis Carranzo) of Ayacucho, while in the CVR list the Shining Path victims are more uniformly spread across the country. More details on the data are available in Appendix A.4.

Now we move to our analysis. Our goal was to estimate the number of killings and disappearances attributable to the State and Shining Path, as well as those that were not identified as either. We used our proposed doubly robust estimator (2.9) with five-fold cross-fitting, and we estimated the q-probabilities via random forests (using the `ranger` package in R). We truncated all q-probability estimates at 0.01. Figure 2.5 shows the estimated number of killings and disappearances along with 95% confidence intervals obtained using the interval (2.12). We estimate the total number of killings and disappearances across groups to be 68,874 (95% CI: 58,543-79,204), close to the estimates in Ball et al. (2003) and the diffuse prior-based estimate in Manrique-Vallier and Ball (2019) (which used an additional list). Overall we find the State responsible for more disappearances, and Shining Path responsible for more killings; however we estimate the number of killings and disappearances by unidentified perpetrators to be larger than that for either group. In terms of the overall killings and disappearances, the estimate for the State are higher compared to the estimate for the Shining Path. We present some more details of the analysis and a location wise estimate comparison for the State and the Shining Path in Appendix A.4.

2.7 Discussion

In this chapter, we study estimation of population size and capture probability in the capture-recapture set-up where two lists are conditionally independent given measured covariates. We make four main contributions to the literature. First, we derive the nonparametric efficiency bound for estimating the capture probability, which indicates the best possible performance of any estimator, in a local asymptotic minimax sense. As far as we know this kind of lower bound result has not appeared in the literature, even in simple settings without covariates. Second, we present a new doubly robust estimator, and study its finite-sample properties; in addition to double robustness, we show that it is near-optimal in a non-asymptotic sense, under relatively mild nonparametric conditions. Third, we give a method for constructing confidence intervals for total population size from generic capture probability estimators, and prove non-asymptotic near-validity. And fourth, we study our methods in simulations, and apply them to estimate the number of killings and disappearances attributable to different groups in Peru during its internal armed conflict between 1980 and 2000.

There are many ways one could extend and build on the work in this chapter. For example, rather than assuming a known pair of lists are conditionally independent given the covariates, one could instead take a sensitivity analysis and/or partial identification approach. For example, one could assume that a pair of lists is only nearly conditionally independent, up to some deviation δ, and estimate bounds on the capture probability and population size accordingly. This relies on weaker assumptions, with the trade-off of yielding less precise inferences. Another extension would be to flexibly estimate conditional capture probability or population size, given a continuous covariate such as age or time. For example, for the internal armed conflict in Peru it might be of interest to estimate the number of victims by age. This would require a non-trivial extension of the current methods, but would be important future work.

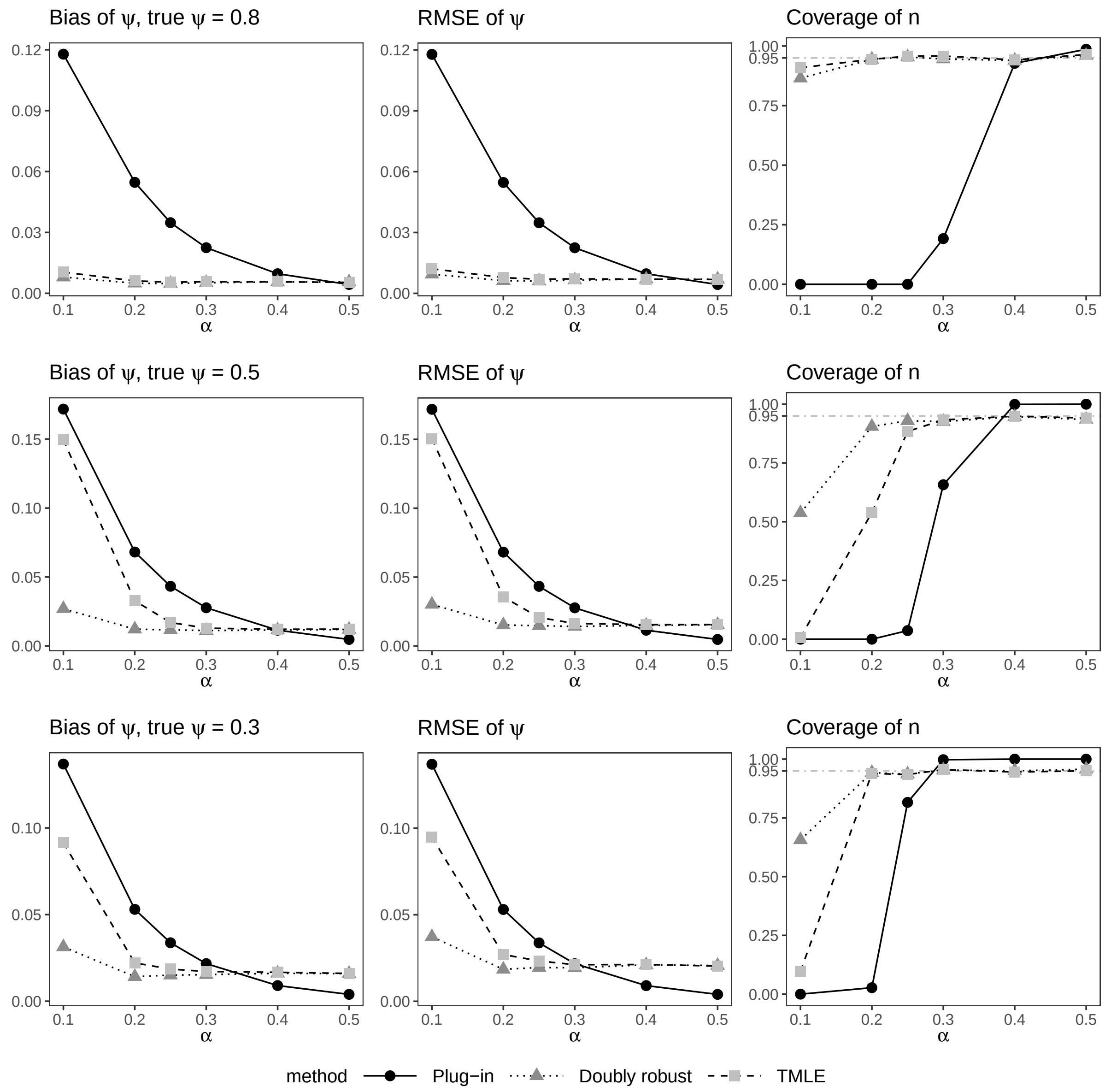

Figure 2.2: Estimated bias, RMSE, and population size coverage, for simulated data with population size $n = 5000$, across true capture probability $\psi \in \{0.8, 0.3\}$, q-probability error rate $n^{-\alpha}$ for $\alpha \in [0.1, 0.5]$, and for three different estimators: the plug-in and two proposed doubly robust estimators.

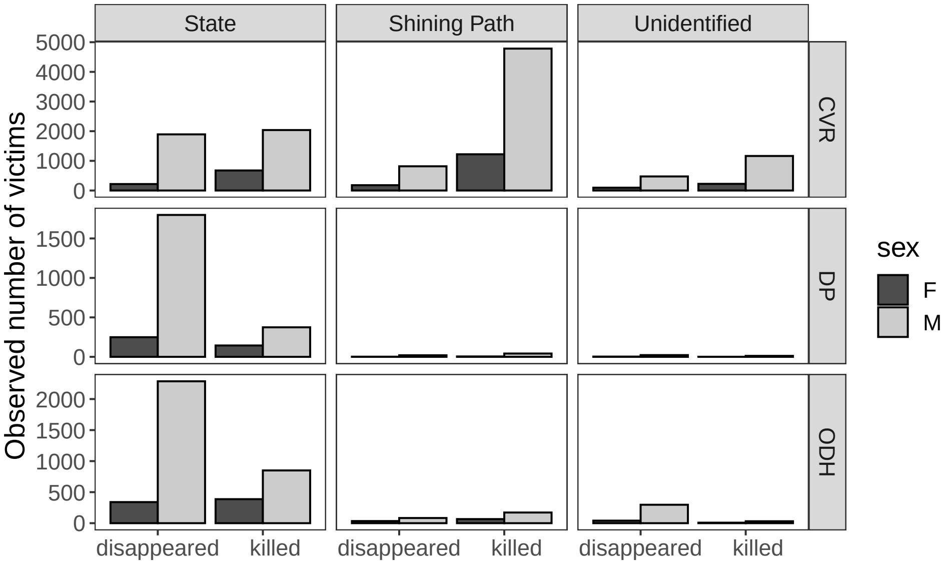

Figure 2.3: This figure shows the observed number of victims for the three lists (CVR, DP, ODH) across the State, the Shining Path and the victims with unidentified perpetrator. Most of the victims are males. The Truth and Reconciliation Commission (CVR) has documented the highest number of victims for the PCP-Shining Path compared to the defender of the People (DP) and the combined NGO's (ODH), whereas the later two sources documented most of the victims for the State and very few (¡70 by DP and ¡400 by ODH) for the Shining Path. Majority of the victims of the State disappeared whereas, most of the victims of the Shining Path were killed.

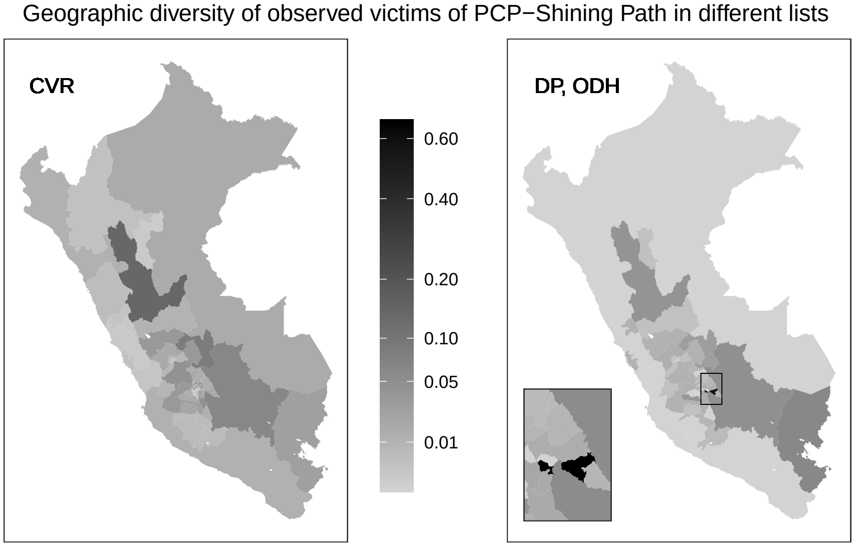
Geographic diversity of observed victims of PCP−Shining Path in different lists
CVR
DP, ODH
0.60
0.40
0.20
0.10
0.05
0.01

Figure 2.4: Geographic diversity of Shining Path victims at strata level, for CVR list (left) and DP and ODH lists (right). The color of each stratum reflects the proportion of all Shining Path victims in the list who were killed or disappeared in that stratum.

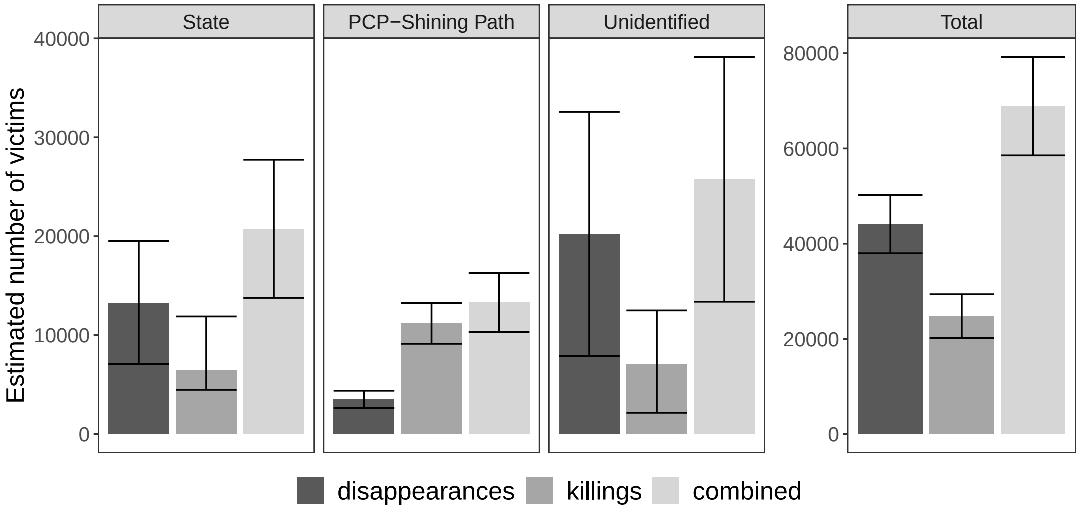

Figure 2.5: Estimated numbers of disappearances and killings (and both together) by perpetrator, as well as total (combined across perpetrators), using the proposed doubly robust method. Bars indicate 95% confidence intervals.

Chapter 3

drpop: Efficient and Doubly Robust Population Size Estimation in R

3.1 Introduction

One crucial step in working with capture-recapture or population size estimation problem, is applying the appropriate identification assumption. Population size estimation is inherently a missing data problem, and hence, one requires some kind of assumption to ensure that the population size is identifiable from the observed data. One should maintain caution while making identifying assumptions, since it can induce bias if not valid for the data (You et al., 2021; Tilling, 2001; Hook and Regal, 1999; Link, 2003; Huggins, 2001). To ensure identifiability, in general, all approaches use some lack of dependence assumption among the lists. The simplest approach works with two lists assuming marginal independence (Petersen, 1896). Some advances in this stream include Schnabel (1938); Darroch (1958); Burnham and Overton (1979) and Lee and Chao (1994). In the presence of covariates, one can use mild assumptions to ensure identifiability. Tilling and Sterne (1999); Huggins (1989); Das et al. (2021) among others assumed that two lists are independent conditional on the covariate and presented non-parametric estimators. This conditional independence assumption is milder than the marginal independence assumption. This assumption can be used for a wide range of data collection scenarios.

And following that, the next step is to account for any heterogeneity present in the data. Real data is often far from homogeneous. Unmodelled or wrongly modelled heterogeneity can also lead to misleading inference (Link, 2003; Carothers, 1973). To account for heterogeneity and/or list dependence, some of the literature used intricate data structures, e.g., complex covariate information. These approaches are mostly model-based. To name a few, there are Link (2003); Carothers (1973); Fienberg (1972); Tilling and Sterne (1999);

Pollock (2002); Huggins (1989); Alho et al. (1993); Yip et al. (2001). Capture probabilities of individuals, i.e., probability of being observed, are often non-linear or complex functions of the covariates (Huggins and Hwang, 2007; Stoklosa and Huggins, 2012) and estimation using linear or strong parametric models might lead to bias. Stoklosa and Huggins (2012) has presented a generalized additive model approach to address this issue.

drpop implements the doubly robust estimators of capture probability and population size from Das et al. (2021), which rely on assuming two lists are only conditionally rather than marginally independent. These methods are flexible yet efficient, with small mean squared error even in non-parametric models involving continuous or high-dimensional covariates.

3.1.1 Existing packages and softwares

There are several R packages and other softwares available for capture-recapture data. Table 3.1 shows a list of some of the existing R packages along with the new drpop. Some of the existing packages are designed for improving estimation and runtime for the classical set-up whereas, others are primarily designed for open population and/or continuous time captures. In the open population set-up, the population is not fixed. There can be addition or deletion. When the population is fixed over the duration of data collection, then it is called a closed population set-up. For this paper, we will focus only on the closed population set-up with discrete capture times. Discrete capture times is the same as a finite number of lists. This set-up generally holds for data collected over a shorter time period.

One of the oldest softwares is MARK (White and Burnham, 1999; White et al., 2001) (extended to R with package RMark by Laake and Rexstad (2008)) and it works on both closed and open population set-ups. For the closed population, it uses the conditional likelihood approach of Huggins (1989, 1991) incorporating individual covariate information. Rcapture (Baillargeon and Rivest, 2007) uses log-linear approach for closed population set-ups implementing the work of Cormack (1989); Rivest and Daigle (2004); Rivest and Baillargeon (2007); Rivest and Lévesque (2001); Cormack (1985); Cormack and Jupp (1991); Frischer et al. (1993). It does not use covariate information but models heterogeneity using lists information. Chao (2014); Chao et al. (2001) presented the R package CARE1 that is designed mainly for closed human populations and uses sample coverage approach. It does not use covariate information either. One of the most recent packages is VGAM (Yee et al., 2015). It is designed for closed population and uses conditional likelihood method while also using covariate information to model heterogeneity. One of the main advantages of VGAM is the ability to model the heterogeneity as non-linear functions of the covariates using vector generalized linear and additive models.

There are other existing softwares and packages, for example, software M-Surge Choquet et al. (2004), and packages like mra (McDonald et al., 2018), marked (Laake et al., 2013), multimark (McClintock, 2015).

These mainly focus on a broader variety of capture-recapture problems, like open population and continuous time captures which are beyond the scope of this paper. For a detailed review and performance comparison, we refer to Bunge (2013) and Yee et al. (2015).

R package	cont. covariate	variance formula	populn. type	param.	nonparam.	eff. & DR
Rcapture 2007			closed/open	✓		
RMark 2008	✓	✓	closed/open	✓		
CARE1 2014			closed		✓	
VGAM 2015	✓	✓	closed	✓		
drpop	✓	✓	closed	✓	✓	✓

Table 3.1: This table lists some R packages for population size estimation. This list is not exhaustive. Our main focus is on the closed population set-up with discrete capture times. We have listed some properties like whether the package can incorporate individual level continuous covariate, has a closed form variance formula, population type it is applicable to, whether it can fit parametric/nonparametric model, and whether it is efficient and doubly robust.

3.1.2 Advantages of drpop

The main goal of drpop is to improve estimation while using complex covariate information to model the heterogeneity. Unlike existing software, the methods in drpop are fully nonparametric, doubly robust, and optimally efficient under weak nonparametric conditions (Das et al., 2021). drpop also lets the user apply their choice of flexible model(s) to capture the heterogeneity in the data. Moreover, it is applicable for data with any number of lists and works with arbitrary discrete or continuous covariates.

In terms of usability, one of the attractions of drpop is that it comes with a lot of options for customization, starting from the model to the level of precision in the estimation. The user can select one or more model(s) for the covariates. The package comes with six in-built models, and is also capable of accepting user-provided model estimates. Further, drpop provides the user with the option to return a baseline estimator and an alternate targeted maximum likelihood estimator (van der Laan and Rubin, 2006) in addition to the proposed doubly robust estimator. In the presence of categorical or numeric discrete covariates, one can also obtain estimates for sub-populations. Other than estimates, there is also an in-built function to simulate data to test models and a plot function for easy inference.

3.1.3 Overview of paper

In this paper, we present the package and some of its applications. Starting in section 3.2, we discuss the data structure for capture-recapture problems and introduce the necessary notations and the identification assumption. In section 3.3, we briefly present the estimation method from Das et al. (2021) to obtain a doubly robust efficient estimator and the formula to obtain a confidence interval. Following this in section 3.4, we present some examples on how to use the drpop for different data types or problems and interpretation of the results. Section 3.5 presents some error rates and performance comparison with some commonly used existing packages to motivate the use of drpop.

3.2 Set-up

In this section, we will discuss the data structure for the capture-recapture data we use. Depending on the approach, there are multiple ways to structure capture-recapture data. In the first subsection, we present our data structure and introduce some of the important notations. In the next subsection, we will discuss the identifiability assumption that the data must satisfy for valid estimates.

3.2.1 Data structure

For a typical capture-recapture problem, the data is a collection of multiple lists. The lists contain information of the capture history of the observed/capture individuals/units. We use K to denote the number of lists. We denote the unknown total population size by n and the number of observed individuals by N. For observed individual i, $i \in \{1, \ldots, N\}$, the capture history is a K-length vector of indicators $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})$. Y_{ik} is 1 if individual i is captured/observed in list k and 0 otherwise. One individual can appear in multiple lists simultaneously, but an observed individual must appear in at least one of the lists i.e. $\mathbf{Y}_i \neq \mathbf{0}$.

In addition to capture profile i.e., lists, we consider the case where we also have covariate information for the observed individuals. We denote the covariate (or covariate vector) for individual i by $\mathbf{X}_i$, which can be used to model the individual-level heterogeneity. We thus denote all data for individual i by $\mathbf{Z}_i = (\mathbf{Y}_i, \mathbf{X}_i)$, and we assume $\mathbf{Z}_i \sim \mathbb{P}$ independently.

The observed data size N is a random draw from the binomial distribution $Binomial(n, \psi)$, where ψ is the capture probability defined by

$$\psi \equiv \mathbb{P}(Y_1 \vee Y_2 \vee \cdots \vee Y_K = 1) = \mathbb{P}(\mathbf{Y} \neq \mathbf{0}).$$

The capture probability ψ is the probability of being observed in at least one of the K lists. By the property of binomial distribution, any estimator for ψ can be transformed to obtain an estimator for n as follows

$$\widehat{n} = N/\widehat{\psi}.$$

However, since we only observe the individuals who satisfy $\mathbf{Y} \neq \mathbf{0}$, we cannot estimate $\mathbb{P}$, and hence, ψ and n directly. Instead, we can estimate the observed data distribution $\mathbb{Q}$, where $\mathbb{Q}$ at a point $\mathbf{z} = (\mathbf{y}, \mathbf{x})$ is defined as

$$\mathbb{Q}(\mathbf{Y} = \mathbf{y},\, \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{Y} = \mathbf{y},\, \mathbf{X} = \mathbf{x} \mid \mathbf{Y} \neq \mathbf{0}) = \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y},\, \mathbf{X} = \mathbf{x})\mathbb{1}(\mathbf{y} \neq \mathbf{0})}{\psi}.$$

If we have d (> 1) dimensional covariates, then the data matrix is of dimension $N \times (K + d)$; each unique individual in its own row. In table 3.2, we present a typical capture-recapture data. The first K columns denote the K lists i.e., data source. The remaining d columns contain the covariate information.

observed individuals	list 1	list 2	...	list K	covariate(s)		
1	Y_{11}	Y_{12}	...	Y_{1K}	X_{11}	...	X_{1d}
2	Y_{21}	Y_{22}	...	Y_{2K}	X_{21}	...	X_{2d}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	Y_{N1}	Y_{N2}	...	Y_{NK}	X_{N1}	...	X_{Nd}

Table 3.2: A typical capture recapture data set from a population with N observed individuals. The data is collected over K sessions or using K sources (lists). Each individual has one or more covariates (d in this example).

3.2.2 Identifiability

As discussed in the previous section, for capture-recapture data, we cannot directly estimate the unconstrained underlying distribution $\mathbb{P}$. Instead, we can estimate the observed data distribution $\mathbb{Q}$. Further, to shift from $\mathbb{Q}$ to $\mathbb{P}$, we need additional assumptions to ensure identifiability. In general, we assume some lack of dependence among the K lists.

The simplest and oldest capture-recapture problems considered only $K = 2$ lists and had no covariates. One can assume that the two lists are independent i.e., $Y_1 \perp\!\!\!\perp Y_2$ to ensure identifiability. This set-up has been used in Petersen (1896). However, the earliest known instance of this approach is by Graunt in the 1600s (Hald, 2003) followed by Laplace (Goudie and Goudie, 2007). It has been further extended to the more than three list case by Darroch (1958) and Schnabel (1938). There have been other modifications to this approach over the years (Jolly and Dickson, 1983; Seber et al., 1982; Bailey, 1952). For more discussion, we refer to Krebs et al. (2014).

Note that it is important that the lists are not completely dependent, i.e., they must have some overlap and they must not be identical, to say the least. Both these cases are uninformative of the unobserved population, and contain the same amount information as the case when we observe only one list. Thus, to ensure identifiability of the total population size, we need some lack of dependence assumption among the lists.

Das et al. (2021) assumes that two lists out of the K lists are collected independently conditioned on the covariate(s). Without loss of generality one can assume that lists 1 and 2 are conditionally independent. One can always reorder the columns to have the two conditionally independent list pair at position 1 and 2.

This assumption has been used very often in past work (Tilling and Sterne, 1999; Sekar and Deming, 1949; Alho et al., 1993; Huggins, 1989; Chao, 1987; Pledger, 2000; Burnham and Overton, 1979; Pollock et al., 1990; Huggins and Hwang, 2007).

Assumption 2. $\mathbb{P}(Y_1 = 1 \mid \mathbf{X} = \mathbf{x}, Y_2 = 1) = \mathbb{P}(Y_1 = 1 \mid \mathbf{X} = \mathbf{x}, Y_2 = 0)$*, where* Y_k *denotes the capture indicator variable for list* k *for* $k = 1, \ldots, K$*.*

This conditional independence assumption is more flexible compared to the conventional marginal independence assumption and accommodates a wide scenario of data collection procedure including the case when the lists have some kind of interaction. For example, when one is collecting data on documented patients at say two different hospitals. Then patients who have already been observed at hospital 1 might have a low probability of being observed again at hospital 2 and vice versa. Hence, the lists of the hospitals are not behaving independently. Now, if we have access to say the location information of the patients, we can describe the behavior of the patients conditioned on that i.e., patients are more likely to visit the hospitals nearer to them. Hence, conditioning on the location, one can assume independence between the two lists.

Another very common identifiability assumption in the capture-recapture literature is the log-linear model introduced by Fienberg (1972). There identifiability is ensured by assuming that the highest order interaction term among all the lists is zero. We refer to Tilling and Sterne (1999); Huggins and Hwang (2011) for more discussion on the differences between identifying assumptions like conditional independence versus log-linear model-based dependence. In particular we refer to You et al. (2021), who give discussion and present methods in a general identification framework without covariates.

In the presence of covariates, we can define the conditional capture probability of an individual by $\gamma(\mathbf{x}) = \mathbb{P}(\mathbf{Y} \neq \mathbf{0} \mid \mathbf{X} = \mathbf{x})$. It is known (e.g., as in Tilling and Sterne, 1999) that under Assumption 2 the capture probability ψ can be identified from the biased observed data distribution $\mathbb{Q}$. Specifically, let

$$\begin{aligned} q_1(\mathbf{x}) &= \mathbb{Q}(Y_1 = 1 \mid \mathbf{X} = \mathbf{x}) \\ q_2(\mathbf{x}) &= \mathbb{Q}(Y_2 = 1 \mid \mathbf{X} = \mathbf{x}) \\ q_{12}(\mathbf{x}) &= \mathbb{Q}(Y_1 = 1, Y_2 = 1 \mid \mathbf{X} = \mathbf{x}) \end{aligned}$$

denote the observational probability (under $\mathbb{Q}$) of appearing on list 1, 2, and both, respectively. These probabilities will be referred to as the q-probabilities at various points throughout. They are also called the nuisance functions or nuisance parameters in this problem set-up and, are crucial in the estimation process.

Now, under Assumption 2, we can define the conditional capture probability and the marginal capture probability as follows

$$\gamma(\mathbf{x}) \equiv \mathbb{P}(\mathbf{Y} \neq \mathbf{0} \mid \mathbf{X} = \mathbf{x}) = \frac{q_{12}(\mathbf{x})}{q_1(\mathbf{x})q_2(\mathbf{x})} \tag{3.1}$$

$$\psi \equiv \mathbb{P}(\mathbf{Y} \neq \mathbf{0}) = \left\{\int \gamma(\mathbf{x})^{-1}\, d\mathbb{Q}(\mathbf{x})\right\}^{-1}. \tag{3.2}$$

Using the expression on the right hand side above, we can directly estimate the capture probability ψ and hence, the total population size by using N/ψ from the observed data. We present the baseline and the proposed method in the following section.

3.3 Methodology

In this section, we discuss a simple plug-in estimator and some of its disadvantages. Following that we discuss our new proposed method in Das et al. (2021) and the ways in which it improves upon the plug-in. Under assumption 2, Das et al. (2021) presents two different estimators for ψ and n: (i) a doubly robust (DR) estimator and (ii) a targeted maximum likelihood estimator (TMLE).

The simplest estimator we can obtain from the expression of ψ in the previous section is based on the plug-in principle, i.e., taking the identifying expression and constructing an estimator by replacing unknown quantities with estimates. The plug-in estimators for the capture-probability ψ and the total population size n are therefore

$$\widehat{\psi}_{PI} = \left\{\sum_{i=1}^{N} \frac{\widehat{q}_1(\mathbf{x}_i)\widehat{q}_2(\mathbf{x}_i)}{\widehat{q}_{12}(\mathbf{x}_i)}\right\}^{-1} \quad \text{and} \quad \widehat{n}_{PI} = \frac{N}{\widehat{\psi}_{PI}},$$

where $\widehat{q}_j$ is the estimated probability value of q_j for $j \in \{1, 2, 12\}$ and $\mathbf{x}_i$ is the covariate value for the observed individual i. In principle, the $\widehat{q}_j$ can be estimated with any parametric (logistic, multinomial logistic) or nonparametric (random forest, gradient boosting) models, though the performance of the plug-in can vary greatly depending on what kind of model is used.

In particular, plug-in estimators typically inherit mean squared errors of the same order as their nuisance parameter estimates $\widehat{q}_j$. This means that when using a plug-in the problem of estimating the one-dimensional capture probability/population size is often made as difficult as estimating the d-dimensional q-probabilities. If one has correct parametric models for these probabilities, this is of little concern, but correct parametric models are hard to come by in practice. When using more flexible methods like random forests or gradient boosting, one would inherit the larger mean squared errors necessarily obtained in nonparametric regression problems.

Beyond the issue of plug-ins having potentially large mean squared errors, in general they also do not come with closed-form variance formulas, which is important for constructing confidence intervals. The bootstrap can be used when parametric models are used to estimate the q-probabilities Tilling and Sterne (1999), but in general the bootstrap fails when more flexible methods are used (e.g., ensembles of high-dimensional regressions).

The proposed doubly robust estimator in Das et al. (2021) uses elements from semiparametric theory (Tsiatis, 2006; Bickel and Ritov, 1988; Kennedy, 2016; van der Laan and Rubin, 2006; van der Vaart, 2002b) to tackle some of the deficiencies of the plug-in estimator discussed above. We discuss more about these properties in the following section.

3.3.1 Proposed Estimators

Das et al. (2021) proposed a doubly robust estimator using semiparametric theory and influence functions. More details on general efficiency theory can be found in Bickel et al. (1993), van der Vaart (2002a), and van der Laan and Robins (2003); reviews can be found in Tsiatis (2006) and Kennedy (2016) among others.

Das et al. (2021) showed that the (uncentered) efficient influence function of the capture probability ψ is given by

$$\phi_i = \frac{1}{\gamma(\mathbf{X}_i)} \left\{ \frac{Y_{1i}}{q_1(\mathbf{X}_i)} + \frac{Y_{2i}}{q_2(\mathbf{X}_i)} - \frac{Y_{1i}Y_{2i}}{q_{12}(\mathbf{X}_i)} \right\},$$

where $\gamma(\mathbf{X}_i) = \frac{q_{12}(\mathbf{X}_i)}{q_1(\mathbf{X}_i)q_2(\mathbf{X}_i)}$ is the conditional capture probability of observation i. The efficient influence function is crucial since (i) its variance acts as a minimax lower bound in nonparametric models (van der Vaart, 2002a), and (ii) it can be used to construct efficient estimators that attain the minimax lower bound. Since the expected value of the efficient influence function is the inverse capture probability ψ^{-1}, Das et al. (2021) proposed the following doubly robust estimators for the capture probability and the total population size n

$$\widehat{\psi}_{DR} = \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\phi}_i \right)^{-1} \quad \text{and} \quad \widehat{n}_{DR} = \frac{N}{\widehat{\psi}_{DR}},$$

where $\widehat{\phi}_i$ is obtained by substituting the estimates of the q-probabilities into ϕ_i. This estimator has some very favorable properties such as: (i) $1/n$-rate mean squared errors, even in flexible non-parametric models, (ii) double robustness, (iii) local asymptotic minimaxity, and (iv) asymptotic normality with finite-sample guarantees. We briefly discuss these properties in this paper, and for more details refer to Das et al. (2021).

As a consequence of efficiency theory, the error in estimation using the proposed estimator is of the order of $1/\sqrt{n}$ even when all three nuisance parameters are estimated flexibly. The formal result states that for any sample size N and error tolerance $\delta > 0$, $|(\widehat{\psi}_{dr}^{-1} - \psi^{-1}) - \mathbb{Q}_N \phi| \leq \delta$ with probability at least

$1 - \left(\frac{1}{\delta^2}\right) \mathbb{E}\left(\widehat{R}_2^2 + \frac{\|\widehat{\phi}-\phi\|^2}{N}\right)$. $\widehat{R}_2$ is a second-order error term given by

$$\begin{aligned}\widehat{R}_2 &= \int \frac{1}{\widehat{q}_{12}} \left\{ \left(q_1 - \widehat{q}_1\right)\left(\widehat{q}_2 - q_2\right) + \left(q_{12} - \widehat{q}_{12}\right)\left(\frac{1}{\gamma} - \frac{1}{\widehat{\gamma}}\right) \right\} d\mathbb{Q} \\ &\leq \left(\frac{1}{\epsilon}\right) \|\widehat{q}_1 - q_1\| \|\widehat{q}_2 - q_2\| + \left(\frac{1}{\epsilon^3}\right) \|\widehat{q}_{12} - q_{12}\| \|\widehat{\gamma} - \gamma\|\end{aligned}$$

The bound on $\widehat{R}_2$ holds as long as $(q_{12} \wedge \widehat{q}_{12}) \geq \epsilon$. If all the nuisance parameters are estimated with error $\sim n^{-1/4}$, i.e., a rate commonly found in nonparametric regression problems, then the error is still bounded above by $C/\sqrt{n}$ for some constant C with probability converging to 1 as n and therefore N increases. The plugin estimator however, does not possess this property and in general inherits the slower rate (e.g., $n^{-1/4}$) from the nonparametric estimation of the q-probabilities.

Another important property of the proposed estimator is the double robustness property presented in corollary 2 in Das et al. (2021). This result follows directly from the formula of the second order error term above $\widehat{R}_2$. This result states that if two out of the four quantities γ, q_1, q_2, and q_{12} have small estimation error, then $\widehat{\psi}_{DR}$ and hence, also $\widehat{n}_{DR}$ will have small estimation error. More specifically, we need one of q_{12} and γ to be estimated with small error and, one of q_1 and q_2 to be estimated with small error. This property is useful when one of the two lists is difficult to estimate or is a complex function of the covariates. More details can be found in Das et al. (2021).

Since the proposed estimator $\widehat{\psi}_{DR}$ is a sample average of the estimated efficient influence functions, it has variance nearly equal to the variance of the estimated efficient influence function divided by N, i.e.,

$$var(\widehat{\psi}_{DR}^{-1}) = \frac{var(\widehat{\phi})}{N}.$$

If σ^2 denotes the population variance of ϕ, then one can estimate $var(\widehat{\psi}_{DR}^{-1})$ by $\widehat{\sigma}^2/N$ where $\widehat{\sigma}$ denotes the estimator of σ. The variance of the efficient influence function divided by N, i.e., σ^2/N, acts as a minimax lower bound in the sense that it is the lowest possible mean squared error any estimator can achieve in a local neighbourhood. The mean squared error of the proposed estimator is close to this bound for a large sample size, e.g., when the nuisance parameters are estimated with errors converging to zero. Das et al. (2021) further presented finite sample analogs of the usual asymptotic minimax arguments and error bounds, including finite-sample distance from a Gaussian distribution.

All the properties we discussed for the capture probability estimate also apply to the total population size estimate. In population size estimation problems, the main interest is often in a confidence interval for n. In the next section, we discuss the properties of the estimated confidence interval.

Confidence interval estimation

One of the main motivations behind using the proposed estimator is that it has a well defined variance formula, as discussed in the previous section. One can estimate $var(\widehat{\psi}_{DR}^{-1})$ by using the unbiased sample variance of $\widehat{\phi}$ scaled by N. This can be used to obtain the variance estimator for the estimated total population size $\widehat{n}_{DR}$ which is given by

$$\widehat{var}(\widehat{n}_{DR}) = N\widehat{var}(\widehat{\phi}) + \frac{N(1-\widehat{\psi}_{DR})}{\widehat{\psi}_{DR}^2},$$

where $\widehat{var}(\widehat{\phi})$ is the unbiased variance estimate of $\widehat{\phi}$. For the derivation, we refer to Das et al. (2021). This variance formula for the estimated total population size can be applied more generally to any estimator that can be approximated by a sample average. The estimated $(1-\alpha)\times 100\%$ confidence interval is

$$\widehat{CI_n} = \widehat{n} \pm z_{\alpha/2}\sqrt{\widehat{var}(\widehat{n}_{DR})}.$$

The finite sample validity/coverage error for this interval is presented in Das et al. (2021). In particular they show the coverage error is bounded above as

$$\left|\mathbb{P}\left(\widehat{\mathrm{CI}_n} \ni n\right) - (1-\alpha)\right| \lesssim n^{(1-4\beta)/2} + \frac{1}{\sqrt{n}},$$

if the nuisance estimators have mean squared errors of order $O(n^{-2\beta})$. Hence, if $\beta > 1/4$, then for any $\epsilon > 0$, there exists an N_ϵ, such that the coverage error is less than ϵ for any $N > N_\epsilon$.

The availability of a closed form formula for the variance and hence, the confidence interval allows for simple inference. Moreover, this also can be used to study the effect of the constituent elements on the variance of the estimate. This eliminates the need to use methods like bootstrap, for example, which can be computationally intensive or not guaranteed to provide valid coverage.

Das et al. (2021) also presented an alternate targeted maximum likelihood estimator $\widehat{\psi}_{TMLE}$ and the associated $\widehat{n}_{TMLE} = N/\widehat{\psi}_{TMLE}$. This estimator has the same properties as $\widehat{\psi}_{DR}$, but the method of calculation uses clever covariates in the targeted maximum likelihood algorithm (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). This estimator does not have a closed form expression. For simplicity, we focus on the original doubly robust estimator in this paper.

3.4 Implementation using drpop

In this section, we illustrate the various functions available in drpop and their implementation in detail. The main goal of drpop is to easily evaluate a doubly robust efficient estimate of the total population size and

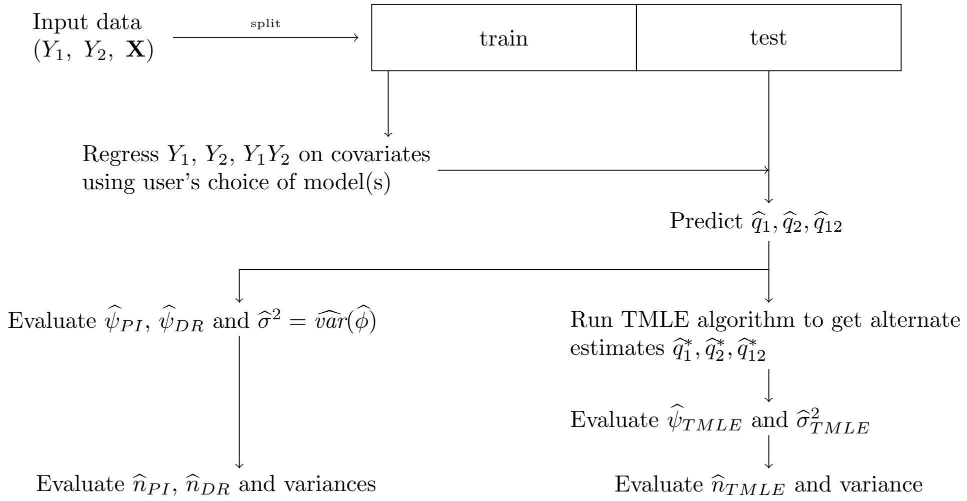
Input data
$(Y_1,\ Y_2,\ \mathbf{X})$
split
train
test
Regress Y_1, Y_2, Y_1Y_2 on covariates using user's choice of model(s)
Predict $\widehat{q}_1, \widehat{q}_2, \widehat{q}_{12}$
Evaluate $\widehat{\psi}_{PI}$, $\widehat{\psi}_{DR}$ and $\widehat{\sigma}^2 = \widehat{var}(\widehat{\phi})$
Run TMLE algorithm to get alternate estimates $\widehat{q}_1^*, \widehat{q}_2^*, \widehat{q}_{12}^*$
Evaluate $\widehat{\psi}_{TMLE}$ and $\widehat{\sigma}^2_{TMLE}$
Evaluate $\widehat{n}_{PI}$, $\widehat{n}_{DR}$ and variances
Evaluate $\widehat{n}_{TMLE}$ and variance

Figure 3.1: The above figure depicts the estimation procedure followed by the estimation function in the package. For simplicity, we show only two lists, and only one train and one test sample. The functions in the package however, uses cross-fitting to utilize the whole observed data.

an associated confidence interval from any capture-recapture data with covariate information. The package is capable of handling high-dimensional and/or complex covariates, both discrete and continuous. It also contains some additional functions that aid in method design, model testing, and inference.

Before diving into the implementation, we discuss the estimation process for a given dataset. In the previous section, we discussed three possible estimators: the plug-in (PI), the proposed doubly robust (DR) and the targeted maximum likelihood estimator (TMLE). drpop has the option to return all three of these estimators, though the default is just to return the doubly robust estimator. To illustrate the steps in the estimation process, we present a flow chart in Figure 3.1 that evaluates the estimates for the capture probability ψ and the total population size n for a capture-recapture dataset with two lists. For the case of more than two lists ($K > 2$), drpop returns the estimates for every possible list-pair unless specified otherwise. Moreover, drpop uses cross-fitting to achieve complete efficiency (Zheng and van der Laan, 2010; Robins et al., 2008; Chetverikov et al., 2021). But, for simplicity, we only present a simple sample-splitting in the flow chart.

Following is the list of functions available in the package along with their brief descriptions.

1. `simuldata`: Generate two or three list toy data with desired features
2. `informat`: Check if data is in format
3. `reformat`: Reorder columns to put data in format
4. `qhat_logit`, `qhat_mlogit`, `qhat_gam`, `qhat_ranger`, `qhat_sl`, `qhat_rangerlogit`: Estimate nuisance parameters q_1, q_2, q_{12}

5. `tmle`: Obtain targeted maximum likelihood estimates of nuisance parameters

6. `popsize`: Estimate population size from raw data or with user provided nuisance estimates

7. `popsize_cond`: Estimate population size from raw data conditional on a discrete covariate

8. `plotci`: Plot the results of `popsize`, or `popsize_cond`.

For a given dataset, one only needs to call either `popsize` or `popsize_cond` to get the estimates of the capture probabilities, total population size, and the confidence intervals.

In this section we briefly describe some data types one can come across and the interpretation of the results. To illustrate the use, we will use toy data examples. A typical dataset in the capture-recapture format has at least two binary columns (corresponding to two or more lists) indicating list-wise capture profiles and one or more covariate column(s). Each observed or captured individual has their own row.

3.4.1 Choice of models for nuisance parameters

The estimation of the population size and the capture probability requires modelling the capture profiles conditional on the covariates. drpop provides six modelling choices listed as follows.

1. `logit`: Fits logistic regression using R function `glm`.

2. `mlogit`: Fits multinomial logistic regression using R function `multinom` in package `nnet`.

3. `gam`: Fits simple generalized additive model from the R package `gam`.

4. `ranger`: Fits random forest model from the R package `ranger`. Suitable for high dimensional covariates.

5. `rangerlogit`: Fits an ensemble of random forest and logistic model.

6. `sl`: Fits different SuperLearner algorithm from the library provided by the user from the R package `SuperLearner`. Returns estimates using a combination of the fitted models. The user can specify the library of models via `sl.lib`.

The computation time varies based on the above models. The parametric models `logit` and `mlogit` are generally the fastest. However, they can lack flexibility, making resulting estimates biased if the nuisance parameters are more complex functions of the covariates. `gam` is slightly slower than the parametric models, but is still comparably fast enough for practical purposes. The flexible nonparametric models `ranger` and `rangerlogit` can be slower to run than these previous models. However, being flexible, these methods can accommodate more complex nuisance functions. `rangerlogit` is the default model in drpop and the performance statistics are presented in section 3.5. `sl` is the slowest depending on the models passed into `sl.lib`. This is because it aggregates multiple models, returning the best estimator combining the individual models using cross-validation. drpop provides the user with the option to parallelize `sl` using `snowSuperLearner` from the R package `SuperLearner`, which is supported on all three of Windows, MacOS and Linux.

For simplicity, we apply some of these models on a toy dataset, `listdata` as shown below. The true population size is 2000 and there are $N = 1610$ rows in the data. The columns `y1`, `y2` and `x1` show list 1 captures, list 2 captures and a continuous covariate. The empirical capture probability is approximately 0.85.

```
> head(listdata, 3)
  y1 y2       x1
1  1  1 2.159287
2  0  1 2.654734
3  1  1 5.338062
```

The function `popsize` returns the estimates via `nuis` for the observed data probabilities q_1, q_2 and q_{12} which are often called the nuisance estimates. It also returns the fold assignment for each row. For simplicity, we use two folds and plot the estimated nuisance parameters.

```
> qhat = popsize(data = listdata, funcname = c("rangerlogit", "logit", "gam", "mlogit", "sl"), nfolds =
```

The dataframe `qhat$nuis` contains the estimates for q_1, q_2 and q_{12} for each model supplied by the user for each row of the data. `qhat$idfold` shows the fold assigned to each row. Figure 3.2 shows the estimated probabilities along with the capture profiles of list 1, list 2 and the two lists simultaneously. One also has the option of using models outside the drpop package and obtain estimates which we will present later in section 3.4.5. Next, we illustrate some examples of application of the package starting from the simplest case.

To ensure that the estimator is valid, we required that all the nuisance parameter estimates, which are probabilities, are bounded away from zero. The default bound is 0.005. One can change this using the argument `margin` in `popsize` or `popsize_cond`.

3.4.2 Two-list case with covariates

The simplest capture-recapture data has two lists with one or more covariates. We present the toy data, `listdata` with true population size 5000 and two continuous covariates.

```
> head(listdata, 3)
  y1 y2       x1        x2
1  1  1 5.342829 0.4682059
```

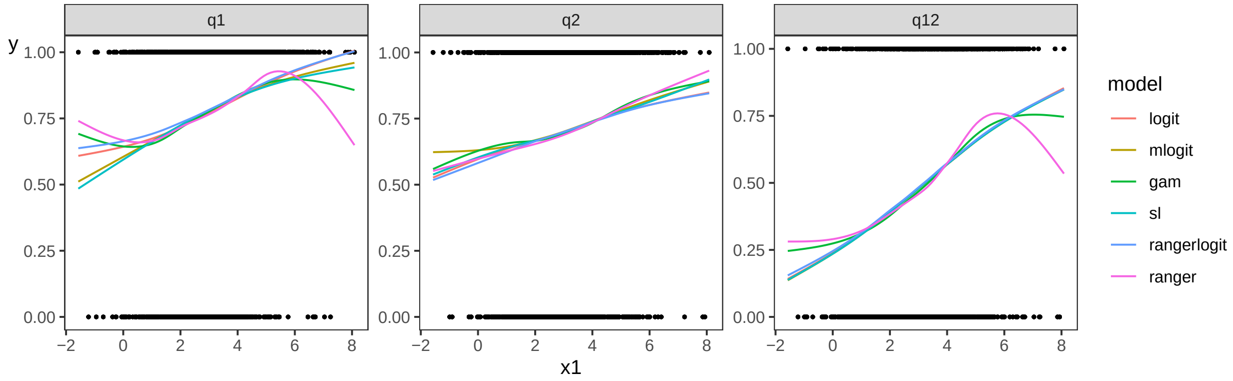

Figure 3.2: The plot shows the smoothed estimated `q1`, `q2` and `q12` for five different models against the scalar covariate `x1`. The points at 0 and 1 show the capture profiles of the individuals i.e., Y_1, Y_2 and Y_1Y_2 respectively.

```
3   1   0 3.700239 2.0279143
4   1   1 4.279882 3.3915513
> result = popsize(data = listdata, funcname = c("logit", "gam", "mlogit", "sl"))
```

To obtain the total population size estimate, we call the function `popsize`. This function accepts the data frame `listdata` as `data` and list of model names, `funcname` which are to be used to estimate the nuisance parameters (q_1, q_2, q_{12}). `popsize` returns a list of objects which include the estimated population size, estimated capture probability, estimated variance and the 95% confidence intervals. Above we print only the estimated capture probabilities `psi`, estimated population sizes `n`, estimated σ `sigma`, estimate standard deviation of $\widehat{n}$ `sigman` and the 95% confidence intervals `cin.l, cin.u` for the total population size. The columns `listpair, model` and `method` indicate the list pairs (lists 1 and 2 in this case), model used to estimate heterogeneity from covariates, and the formula for estimation of the target parameters ψ and n respectively.

Remark 18. *Setting arguments `PLUGIN` and `TMLE` to `FALSE` will return only the `DR` (proposed doubly robust) estimates. We also plot the confidence intervals using the `plotci` function.*

```
> result = popsize(data = listdata, funcname = c("gam", "logit",
           "mlogit", "sl"), PLUGIN = TRUE, TMLE = TRUE)
> print(result)
   listpair  model method   psi sigma    n  sigman cin.l cin.u
1       1,2    gam     DR 0.910 0.440 4978  37.032  4905  5051
2       1,2    gam     PI 0.917 0.440 4941  36.433  4870  5012
3       1,2    gam   TMLE 0.912 0.595 4968  45.649  4878  5057
```

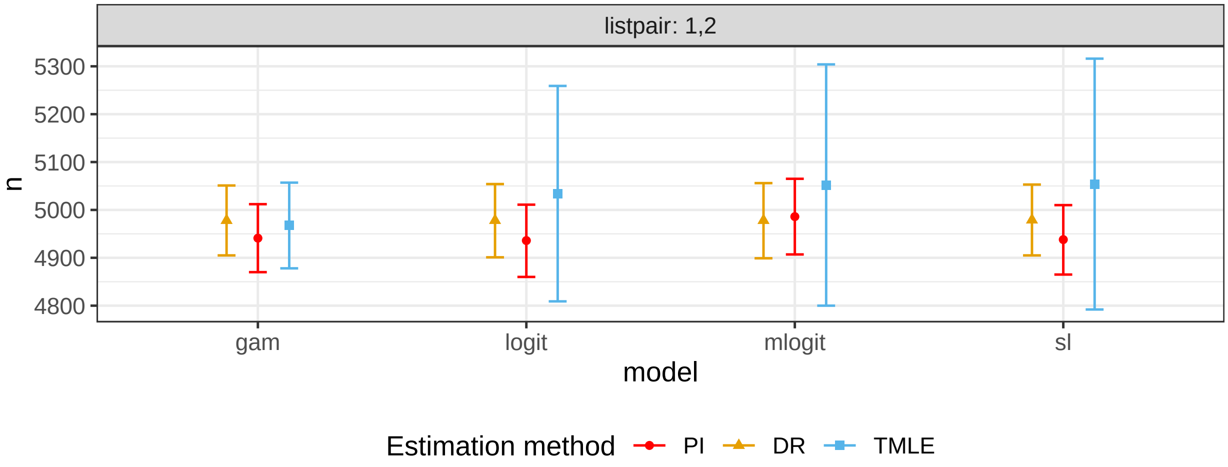

Figure 3.3: The above plot shows the estimated confidence interval for n for different models. The true population size is 5000. The term list-pair specifies the two lists used for the estimation.

```
4       1,2  logit     DR 0.910 0.478 4978  39.073  4901  5054
5       1,2  logit     PI 0.918 0.478 4936  38.423  4860  5011
6       1,2  logit   TMLE 0.900 1.670 5034 114.852  4809  5259
7       1,2 mlogit     DR 0.910 0.498 4978  40.184  4899  5056
8       1,2 mlogit     PI 0.908 0.498 4986  40.311  4907  5065
9       1,2 mlogit   TMLE 0.897 1.875 5052 128.443  4800  5304
10      1,2     sl     DR 0.910 0.452 4979  37.689  4905  5053
11      1,2     sl     PI 0.917 0.452 4938  37.034  4865  5010
12      1,2     sl   TMLE 0.896 1.954 5054 133.735  4792  5316
> plotci(result)
```

Remark 19. *Since the plug-in estimator has no known variance formula, we use the same variance formula as the proposed estimator for the calculation of the variance of the plug-in estimators.*

3.4.3 Two-list case with conditional estimates

When one has a discrete or categorical covariate in addition to other covariates, it is often of interest to estimate the total population size conditioned on that categorical covariate, i.e., for sub-populations. For example, suppose one has a population of patients in a city and their age, demographic information, and ethnicity as the covariates. Then it can be of interest to obtain the estimated population size for the different ethnicities separately.

We again use a simulated toy dataset to illustrate the implementation. The data has three continuous covariates (`x1, x2, x3`) and one categorical covariate column called `catcov`. `catcov` takes three possible values ‘a’, ‘b’, ‘c’ with equal probability. Total population size is 6000 and each of ‘a’, ‘b’ and ‘c’ appear roughly 2000 times in the whole population. We present the first three rows below.

```
> head(listdata, 3)
  y1 y2       x1       x2        x3 catcov
1  1  1 2.159287 5.897364 3.4173336      b
2  1  0 2.654734 2.075288 0.5961934      a
3  1  0 5.338062 2.156149 2.5186507      c
```

The interest here is to obtain population size estimates conditioned on the categorical variable `catcov`, i.e., for sub-populations with `catcov` value ‘a’, ‘b’ and ‘c’ separately. The function `popsize_cond` is similar to the function `popsize` but returns the result separately for each level of the categorical variable. We specify the categorical covariate to be used for conditioning using the argument `condvar`. To obtain an overall estimate one can use `popsize` as in the previous example.

```
> result = popsize_cond(data = listdata, condvar = 'catcov', funcname = c("mlogit", "gam"), PLUGIN = TRU
> print(result)
   listpair  model method   psi  sigma    n  sigman cin.l cin.u condvar
        1,2 mlogit     DR 0.560  4.821 3040 204.818  2639  3442       b
        1,2 mlogit     PI 0.575  4.821 2960 204.323  2560  3361       b
        1,2 mlogit   TMLE 0.541  7.050 3147 295.398  2568  3726       b
        1,2     sl     DR 0.627  5.501 2715 230.476  2263  3167       b
        1,2     sl     PI 0.606  5.501 2808 230.926  2355  3260       b
        1,2     sl   TMLE 0.637  3.296 2670 141.464  2393  2948       b
        1,2 mlogit     DR 0.596  4.683 3306 213.115  2888  3724       a
        1,2 mlogit     PI 0.590  4.683 3338 213.290  2920  3756       a
        1,2 mlogit   TMLE 0.630  3.401 3126 156.890  2818  3433       a
        1,2     sl     DR 0.612  3.731 3216 171.610  2880  3553       a
        1,2     sl     PI 0.630  3.731 3125 171.019  2790  3460       a
        1,2     sl   TMLE 0.594  5.777 3313 260.696  2802  3824       a
        1,2 mlogit     DR 0.533  7.068 3082 291.147  2511  3652       c
        1,2 mlogit     PI 0.558  7.068 2946 290.526  2377  3516       c
        1,2 mlogit   TMLE 0.489 10.859 3359 444.129  2488  4229       c
        1,2     sl     DR 0.524  6.486 3138 268.283  2612  3664       c
```

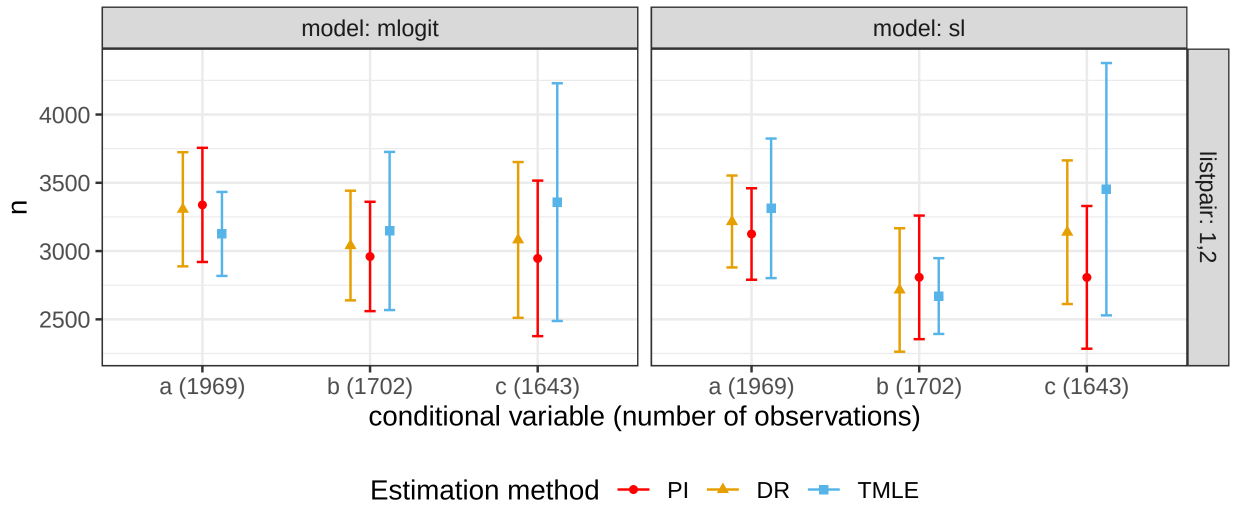

Figure 3.4: The above figure shows the confidence interval for n for three sub-populations and models gam, logit and sl. The sub-populations are obtained from the original population using the values of the `catcov` covariate.

```
        1,2     sl     PI 0.585  6.486 2807 266.666  2285  3330       c
        1,2     sl   TMLE 0.476 11.535 3453 471.596  2529  4377       c
> plotci(result)
```

The result of `popsize_cond` is in a similar format to `popsize`, but it specifies the level of the categorical covariate i.e., the sub-population in a separate column.

3.4.4 Three or more lists

The approach used by drpop assumes that there are two lists which are known to be conditionally independent. However, capture-recapture datasets can often consist of more than two lists. If the analyzer knows the list-pair that is conditionally independent, they can use the functions `popsize` and `popsize_cond` by removing the remaining list columns or by specifying the two list columns to be used for estimation. However, when the analyzer is not aware of the list-pair, the entire dataset can be passed into the estimation functions. drpop returns an estimate for each possible list-pair.

The toy dataset has three list columns as shown below. Now, since we pass more than two list columns into the functions, the output will have the result for the different list-pairs `(1,2)`, `(1,3)` and `(2,3)`.

```
> head(listdata,3)
  y1 y2 y3       x1       x2        x3       x4
1  0  0  1 1.189401 6.737728 0.8531169 1.508898
```

```
2  1  0  1 3.416144 3.079832 3.1891693 4.082209
3  1  0  0 4.626662 3.684374 3.6552886 2.694606
```

For more than two lists, we need to specify the number of list columns using `K` in `popsize`. For simplicity, we evaluate only the doubly robust estimators in this example and exclude the TMLE and the plug-in estimates. The `listpair` column in the result below specifies the list-pair used for the estimation. For example, assuming $Y_1 \perp\!\!\!\perp Y_2$ under the `rangerlogit`, we get $\widehat{n}_{DR} = 29,711$, and assuming $Y_1 \perp\!\!\!\perp Y_3$ under the `rangerlogit` model, we get $\widehat{n}_{DR} = 30,423$.

```
> result = popsize(data = listdata, K = 3, funcname = c("mlogit", "gam", "rangerlogit"), nfolds = 2)
> result
   listpair       model method   psi sigma     n  sigman cin.l cin.u
1       1,2         gam     DR 0.872 0.983 29752 171.595 29416 30089
4       1,2      mlogit     DR 0.873 1.497 29723 249.911 29233 30213
7       1,2 rangerlogit     DR 0.874 1.453 29711 243.040 29235 30187
10      1,3         gam     DR 0.860 1.565 30192 261.799 29679 30705
13      1,3      mlogit     DR 0.851 2.271 30502 373.116 29771 31233
16      1,3 rangerlogit     DR 0.853 2.137 30423 351.831 29734 31113
19      2,3         gam     DR 0.859 2.467 30236 403.749 29445 31027
22      2,3      mlogit     DR 0.859 3.452 30234 560.642 29135 31333
25      2,3 rangerlogit     DR 0.853 2.871 30449 468.210 29531 31367
> plotci(result)
```

The plot function in the package shows the estimated confidence interval for n in Figure 3.5. We note that confidence intervals are relatively shorter for list-pair (1,2) and (1,3). The reason being that the overlap between the lists is larger for (1,2) and (1,3) compared to (2,3). As already discussed previously in section 3.2, the overlap between the conditionally independent lists must be bounded away from 0 and N for better estimation.

If the analyzer is aware of the list pair, then the dataset can be passed into the estimation functions by removing all other list columns or by specifying the list pair. Suppose that the two conditionally independent list columns are `y1` and `y2`. Then the user can either remove column `y3` and pass the data into `popsize`, or he can specify the pair. We illustrate both these approaches below.

```
> result = popsize(data = subset(listdata, select = -c(y3)))
> result = popsize(data = listdata, j = 1, k = 2, K = 3)
```

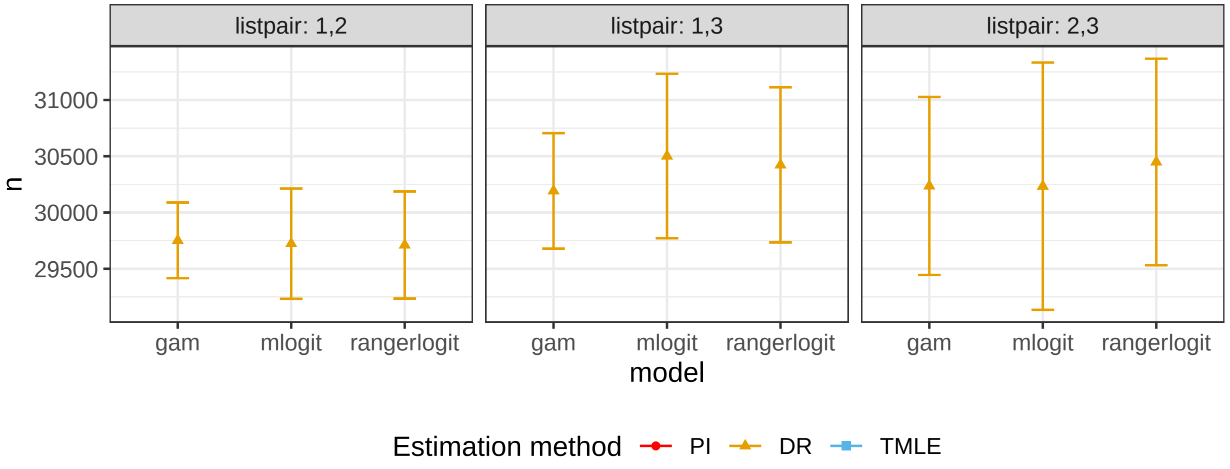

Figure 3.5: The above plot shows the estimated confidence interval for n for three different possible list-pairs under different models. The result for list-pair (1,2) produces narrower intervals closer to the true value 30,000.

3.4.5 Estimation with user provided nuisance estimates

The main purpose of this example is to illustrate how to pass nuisance parameter estimates into `popsize`. This is useful when the user has some background information that suggests modelling the heterogeneity differently than what is available in the package. For simplicity, we illustrate this by passing the nuisance parameter estimates, `nuis` from the output of `popsize` with the default model `rangerlogit`. The toy dataset used has total population size 5000 with two continuous covariates. We show the first few rows of the estimated nuisance parameters in `estim$nuis`. The columns specify the model name (`rangerlogit` in this case) and also the q-probabilities. For more than one model, `estim$nuis` will contain additional columns in the same format. `estim$idfold` specifies the fold assignment for each row. Rows 1 and 2 are assigned to folds 5 and 1 respectively in this example. There are total five folds because `nfolds = 5`.

```
> listdata = simuldata(n = 5000, l = 2, ep = -3)$data
> head(listdata, 3)
  y1 y2       x1       x2
1  1  0 2.159287 2.258739
2  0  1 2.654734 4.691390
3  0  1 5.338062 1.279576
> estim = popsize(data = listdata, funcname = c("rangerlogit"), nfolds = 5)
> head(estim$nuis)
  listpair rangerlogit.q12 rangerlogit.q1 rangerlogit.q2
1      1,2       0.1284399      0.7116891      0.4167508
```

```
2      1,2      0.2425309      0.7532815      0.4892493
3      1,2      0.2012161      0.8156479      0.3855682
4      1,2      0.2832612      0.8827187      0.4005426
5      1,2      0.4362669      0.8679311      0.5683358
6      1,2      0.2755165      0.8208690      0.4546476
> head(estim$idfold)
[1] 5 1 2 2 3 3
```

Once we have the nuisance parameter estimates, we can pass it into `popsize`. As mentioned above, there are multiple ways of executing this. We illustrate the most straightforward approach below by passing `estim$nuis` and `estim$idfold` directly. The result is shown below and presented in Figure 3.6.

```
> result = popsize(data = listdata, getnuis = estim$nuis, idfold = estim$idfold)


>result
  listpair       model method   psi  sigma    n  sigman cin.l cin.u
1      1,2 rangerlogit     DR 0.500  3.355 5045 182.890  4686  5403
2      1,2 rangerlogit     PI 0.593  3.355 4255 176.999  3908  4602
3      1,2 rangerlogit   TMLE 0.402 12.549 6284 637.844  5034  7535
> plotci(result)
```

Remark 20. *In the current version of the package, one can pass nuisance parameter estimates only for one list-pair at a time. The lists can be specified using* `j` *and* `k`*. The default is* `j = 1` *and* `k = 2`*.*

Remark 21. *All the datasets used in the examples in this section are simulated data.*

The package has an in-built function `simuldata` to generate a toy dataset with two or three lists. It can be used to test models by comparing against the true value. The `simuldata` function takes in the number of lists (`K`, default value 2), the number of continuous covariates (`l`), the logical option to include one categorical column (`categorical`, default value `FALSE`) and a numeric parameter to control the capture probabilities (`ep`, default value 0). It returns the empirical capture probability (`psi0`), the simulated dataset (`data`), the simulated dataset with transformed continuous covariates (`data_xstar`) and the list wise capture probability functions (`pi1`, `pi2`, `pi3`) depending on `K`. For example, the function `pi1` returns the probability of being observed in list 1 for the covariate vector passed into it. The dataset with transformed covariates, `data_xstar` can be used to study the robustness of a model.

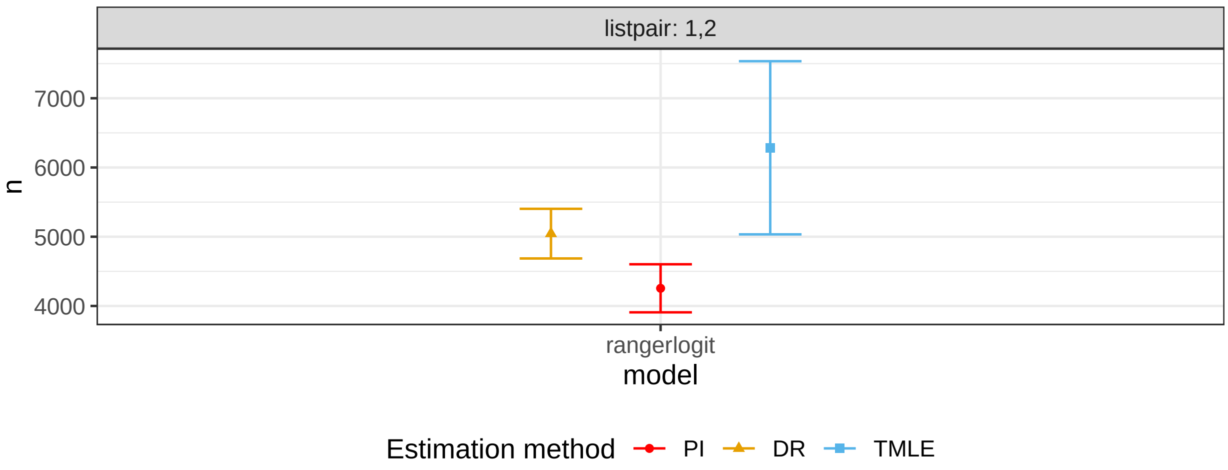

Figure 3.6: The plot above, generated by `plotci`, shows the estimated confidence interval for n for the user provided nuisance estimates under two models, gam and logit. The variable list-pair (1,2) presents the conditionally independent lists. `popsize` returns results for only one list-pair which is the first list pair unless specified otherwise by the user.

3.5 Performance

To motivate the use of the doubly-robust estimators of drpop for closed population, we present some summary statistics of its performance in a simulated set-up. The main focus of drpop is to flexibly estimate the total population size and at the same time to achieve optimal $1/n$ mean squared errors. The identifiability assumption used in drpop requires just two lists which are independent conditional on the covariates. We use simulated data that roughly satisfies this assumption to measure the performance of the proposed method. First, we present a comparison of the proposed doubly robust estimator in drpop against the baseline plug-in estimator under the flexible nonparametric set-up and also when any of the covariates are not correctly specified. Following that, we present some performance comparison of the closed population set-ups of the packages Rcapture, CARE1, VGAM and drpop.

3.5.1 Performance in simulated set-up

In this section, we evaluate the performance of the proposed method in the package over a 100 iterations and different simulation set-ups. Our simulation set-up ensures that the two lists are independent conditional on the covariates. The goal is to compare the performance against the baseline plug-in estimator. Moreover, we also compare the robustness of the estimators when the covariates are not correctly specified or are transformed. For the later case, flexible non-parametric models would prove useful. drpop has the choice of several such models. But for this section, we only use the default model `rangerlogit` which is an ensemble of logit and random forest models.

We simulate a data-frame using the `simuldata` function for two lists. The true population size is 5000 for each iteration. We generate data with true capture probabilities 0.36 and 0.75 separately. One can set the parameter `ep` equal to -2.5 and -1 respectively for the same. The code used to generate the data is

```
> datalist = simuldata(n = 5000, l = 1, ep = -2.5)
```

where `n` is the true population size and `l` is the number of continuous covariates. The default number of lists in two. One can access the simulated data using `datalist$data` and `listdata$data_xstar` where the later data-frame contains transformed (misspecified covariates).

We evaluated the bias, the RMSE (root-mean-square-error) and the empirical coverage by

$$\widehat{bias} = \frac{1}{100}\sum_{i=1}^{100} |\widehat{n}_i - 5000|^2, \quad \widehat{RMSE} = \sqrt{\frac{1}{100}\sum_{i=1}^{100} (\widehat{n}_i - 5000)^2},$$

$$\text{and } \widehat{coverage} = \frac{1}{100}\sum_{i=1}^{100} \mathbb{1}\left(\widehat{cin.l}_i \leq 5000 \leq \widehat{cin.u}_i\right),$$

where i is the iteration, $\widehat{n}_i$ is the estimated population size at iteration i, and $\widehat{cin.l}_i$ ($\widehat{cin.u}$) denote the estimated lower (upper) limit of the 95% confidence interval. At each iteration, we generate a dataset independently of the other iterations. We evaluate these three quantities for both the capture probabilities, and also under the correct covariate data and misspecified covariate data. The results are shown in Figure 3.7. Overall, both methods perform better when we have a higher capture probability or the correct covariates. The proposed method (`DR`) has lower bias, lower RMSE and a higher empirical coverage compared to the plug-in `PI` estimator under both correct covariates and mis-specified covariates.

3.5.2 Comparison to other packages

We present some comparisons with existing R packages that can work for closed populations. We use the functions `closedp`, `estN` and `vglm` from the packages Rcapture, CARE1 and VGAM respectively. We have considered two set-ups: two list case and three list case. CARE1 requires more than two lists for its sample coverage approach and hence, we drop this package in the two list case. Rcapture uses log-linear models as discussed in Baillargeon and Rivest (2007) and does not use covariate information. It is designed to use information from many lists to model the heterogeneity. VGAM uses log-likelihood approach and can incorporate continuous covariate information using generalized linear/additive models. We used simulated data to compare the performance of drpop against the models in the packages Rcapture, CARE1, and VGAM in a closed population set-up in the following two subsections. We note that these packages are based on assumptions different from ours. We present the comparison result for the sake of completeness.

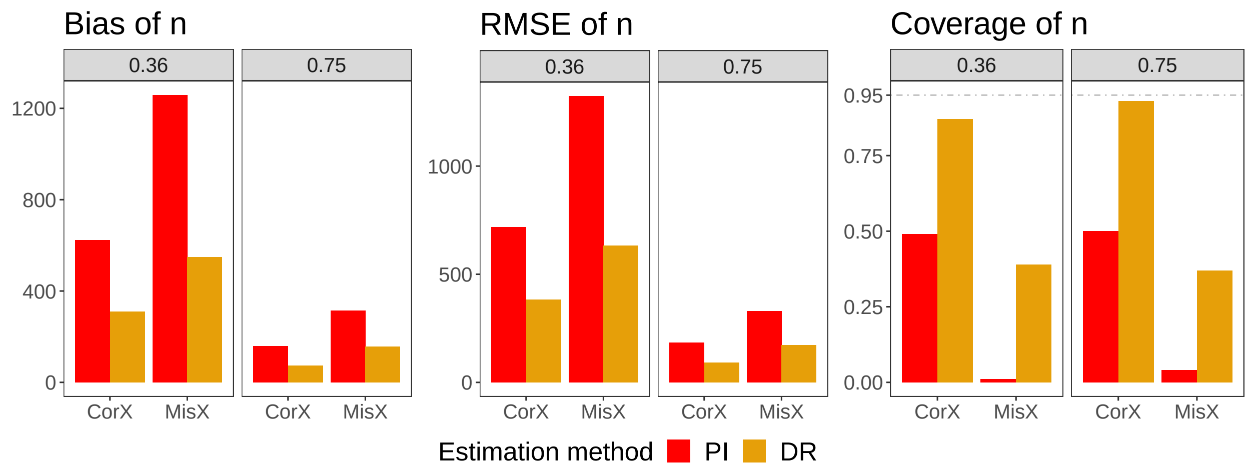

Figure 3.7: This figure shows the estimated average bias, root-mean-square-error (RMSE) and the empirical coverage in the estimation of the total population size `n`. The two facets show the true capture probability which is also the (×100)% of the population observed. `CorX` and `MisX` refer to estimation using data with original covariates and transformed(mis-specified) covariates respectively. The doubly robust (`DR`) estimator has better performance in the set-up shown.

Two list case

We begin with the simple case where we have only $K = 2$ lists with some covariate information. The data is simulated using the simuldata function with parameters `K=2, l=1, ep=-1.5` i.e., the covariate is of dimension one. The total population size takes values in (3,000, 6,000, 9,000, 12,000, 15,000) and the true capture probability is approximately 0.63.

Rcapture function `closedp` only fits three models (`M0` for no henerogeneity, `Mt` for list heterogeneity and `Mb` for heterogeneity based on first capture) when there are only two lists. For a full list of models, one can refer to Baillargeon and Rivest (2007). For `vglm` from package VGAM, we used posbernoulli.t to include list and individual heterogeneity. For drpop, we used the `rangerlogit` model to calculate the doubly robust estimator. Both drpop and VGAM use covariate information. Hence, we further compare their performance in terms of robustness of errors in covariate information i.e., transformed covariates. We applied them on data with the correctly specified/original covariates, and then on data with transformed/mis-specified covariates as in Figure 3.7. The results are presented in Figure 3.8.

We removed the estimates of the `Mb` model from the plot because, it had significantly large errors compared to the other methods (this is expected based on the simulation set-up). In the above two list set-up, the estimate using the drpop and VGAM packages have bias and RMSE decreasing with the total population size at a faster rate compared to models `M0` and `Mt`. The coverage of the estimated confidence intervals is also closer to the nominal level of 95% when the covariates are correctly specified. VGAM has slightly better coverage when the covariates are correctly specified. This is a consequence of the simulation set-up where

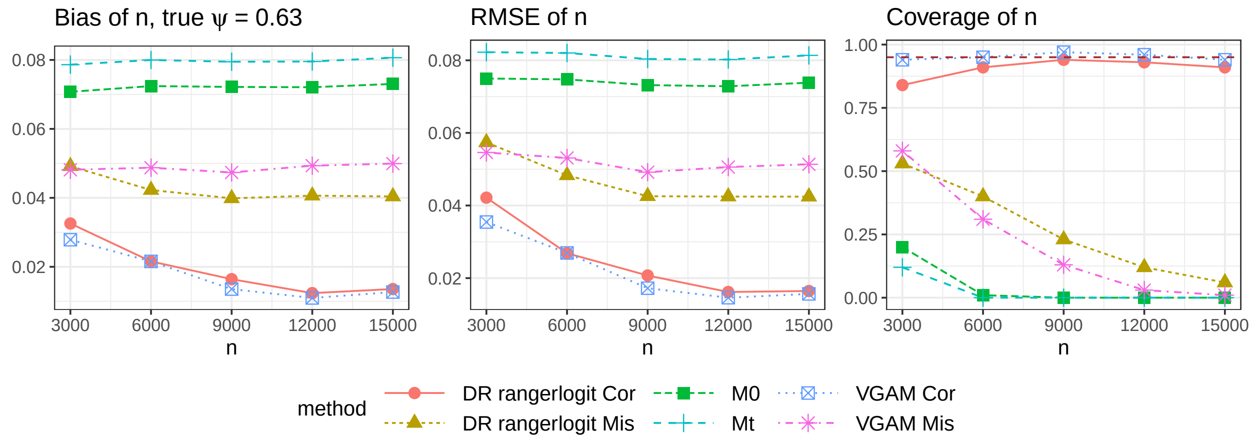

Figure 3.8: The absolute bias, root mean square error (RMSE) scaled by the true n, and empirical coverage of n from the default model in drpop (`DR rangerlogit`), the `M0` and `Mt` models from Rcapture and the model in VGAM. The true population sizes are shown on the x-axis and the the true capture probability is 0.63, i.e., we are observing around 63% of the population. For drpop and VGAM we present results with correctly specified covariates (Cor) and transformed/mis-specified covariates (Mis).

the actual list probabilities are additive functions of the covariates. However, for the mis-specified covariates, drpop has slightly better performance for larger sample sizes.

Remark 22. *The performance result in Figure 3.8 is not necessarily a general phenomenon. This can change based on the simulation set-up, for example. More exploration is needed to figure out if this is general.*

Three list case

In this section, we apply our method and functions from the three packages in a three list set-up ($K = 3$). Our goal in this section is to show that the performance of drpop with the default parameter values, at least matches the performance of Rcapture, CARE1 and VGAM. We again note that these packages are developed based on assumptions different than those of package drpop. Hence, we do not expect unbiased estimates.

We use `simuldata` function to generate toy population with three lists and three dimensional continuous covariates. We set `ep` at -5 and -3 to get true capture probability `psi0` equal to 0.34 and 0.80 respectively. We set the true total population size at 15,000 and 5,000 for 0.34 and 0.80 respectively, since a low capture probability requires a larger number of observations for good estimation. The number of observations for the two set-ups are approximately 5,100 and 4,000 for each iteration. For the above set-up we generated a simulated dataset 100 times and estimated the population size for each.

To compare the performance, we present the boxplot of the scaled bias $(\widehat{n} - n)/n$ of each iteration for the different models in Figure 3.9. The estimation models from the four packages are marked by colors. The doubly robust estimator is `DR rangerlogit` using only the first two lists for simplicity. For Rcapture, we excluded the `Mb` and `Mbh` models because they have large error which is expected under the current

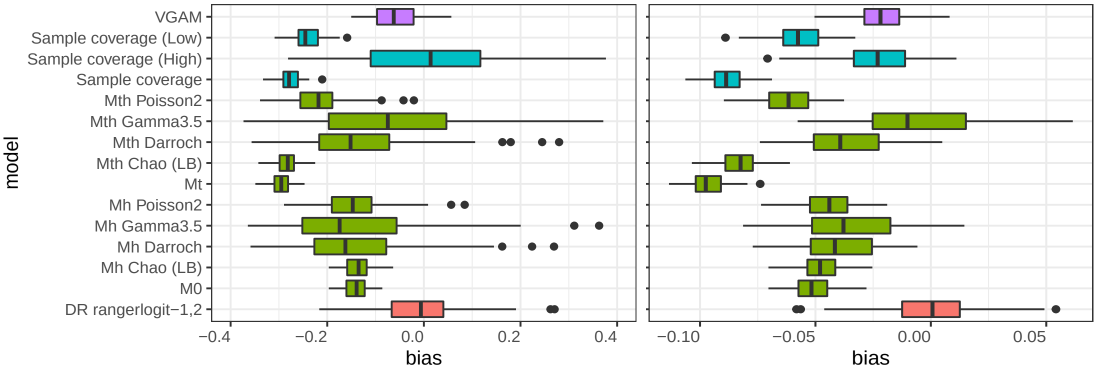

Figure 3.9: Scaled bias $(\widehat{n}-n)/n$ in the estimation of the total population size using four different packages: violet (VGAM), teal (CARE1), green (Rcapture) and red (drpop). CARE1 and Rcapture return multiple estimates. On the left we observe 34% of the whole data, and on the right we observe 80% of the data.

simulation set-up. All the estimators display better performance (lower bias and/or lower variance) for capture probability 0.8 compared to 0.34. For the specific set-up used with 80% observed data, the proposed method and `Mth Gamma3.5` have bias closest to 0 followed by `VGAM` and `Sample coverage (High)`. Whereas, for the 34% observed data set-up, the proposed method and `Sample coverage (High)` have bias closest to 0 followed by `VGAM` and method `Mth Gamma3.5`.

Remark 23. *The performance result in Figure 3.9 is not necessarily a general phenomenon. More exploration is needed to figure out if that is the case.*

Summarizing the results and the advantages of drpop, it is capable of incorporating high dimensional and complex covariates as well as interaction among the covariates. The user can choose from several flexible modelling options that are provided in the package or also, use their own models to estimate the nuisance parameters. Under the identification assumption of conditional independence between two lists, the proposed estimator in drpop also handles mis-specified covariates better compared to the naive plug-in estimator. Further, attributed to the bias-correction step, the estimation under small capture probability (small observed sample) is also better compared to the plug-in estimator.

3.6 Discussion

In this paper, we have presented the R package drpop to implement a new doubly robust estimator of the total population size and an associated confidence interval from incomplete lists. The package provides users with many choices for flexibly modelling the heterogeneity which usually exists in real data. Further, the proposed method implemented in the package (Das et al., 2021) exploits efficiency theory so that it achieves beneficial properties such as (i) $1/n$ mean squared errors even in flexible nonparametric models, (ii) double robustness, (iii) minimax optimality, and (iv) near finite-sample normality.

One of the main advantages of drpop is that it can model the heterogeneity in the data as complex functions of discrete and/or continuous covariates. This is useful when the capture probabilities (nuisance parameters i.e. q_1, q_2, q_{12}) of the individuals do not depend linearly on the covariates. More discussion on this can be found in Yee and Mitchell (1991); Crawley (1993); Gimenez et al. (2006); Bolker (2008); Schluter (1988), and Yee et al. (2015). Yee et al. (2015) also created an R package VGAM which addresses this issue via vector generalized models. The availability of flexible models in drpop makes it easy for users to obtain good estimates for such datasets as well. The users also have the option to fit their own models to estimate nuisance parameters and pass them into the package functions to obtain total population size estimate and confidence interval(s).

The estimation method implemented in drpop exploits modern advances in nonparametric efficiency theory. This ensures that even when one is using flexible nonparametric methods, the rate of convergence (i.e., mean squared error) is not compromised. Typically, plug-in estimators inherit convergence rates from the estimators of the more complex nuisance parameters like q-probabilities. However, because of the form of the proposed estimator, we can still achieve $1/n$ mean squared errors, even when the nuisance functions are estimated flexibly at slower rates. Further, the estimate is doubly robust against errors in the estimation of the nuisance parameters. In particular, even when either one of q_1 and q_2 is estimated with large errors, or one of q_{12} and γ is estimated with large errors, the $\widehat{\psi}$ and n still have bounded errors as long as q_{12} and $\widehat{q}_{12}$ are bounded away from zero. More details and explanation can be found in Das et al. (2021). Further, as a consequence of efficiency theory, this estimator is near minimax optimal in finite samples and has a nearly normal distribution, permitting simple but valid confidence interval construction.

We have presented some simulation results in section 3.5 to show the advantages of the proposed estimator against the baseline method. We have also provided some simulation results to compare the performance of the proposed method in drpop against some of the existing widely used R packages for the closed population set-up. Our goal is to show that when the capture probability depends on covariates and when our mild identifiability assumption holds, the performance of drpop is reliable and comparable to some of the existing methods for the given set-up.

Alongside the proposed doubly robust estimator, this package also provides the user with the choice of evaluating the baseline plug-in estimator and an alternate targeted maximum likelihood estimator (TMLE). Some of the other functions this package can perform are (i) simulate toy data for model training and study design, (ii) estimate total population size and other parameters and other information for sub-populations based on a categorical covariate, and (iii) plot the results with an in-built function for easy and fast interpretation. A full list is presented in section 3.4.

Chapter 4

Nonparametric estimation of population size from conditional capture-recapture designs under partial identification

4.1 Introduction

Population size estimation in an important problem in many areas of sciences. Capture-recapture design denotes data consisting of two or more lists from the population (Petersen, 1896; Chao, 1987; Otis et al., 1978). This problem set-up requires additional assumptions on the lists to ensure identification of the parameters. Some commonly used assumptions are list independence, log-linear models and conditional independence between lists. You et al. (2021) has presented estimators under various identification assumption. For real data, seldom does one know whether the assumptions are satisfied. In recent literature, estimation under relaxed assumptions has gained traction. For example, Chan et al. (2020) presented log-likelihood models when some sources have very little or no overlaps to reflect on the chosen assumption. This motivates us to explore a more relaxed set-up of Das et al. (2021).

Das et al. (2021) has presented efficient and doubly robust estimators under the assumption that two lists are independent conditioned on individual covariate information. This assumption has been extensively studied in the capture-recapture literature, for example in Sekar and Deming (1949); Pledger (2000); Pollock et al. (1990); Tilling and Sterne (1999); Huggins (1989); Chao (1987); Alho (1990). For real data, the knowledge of the data collection procedure is not always available. In such a scenario, it is hard to ensure the validity of the assumption(s) on the lists. We extend this assumption to explore the case where only partial identification is possible. In this paper, we extend the approach of Das et al. (2021) to a more relaxed set-up when the two lists can deviate from the conditional independence assumption. This approach is applicable in the presence of two or more lists, but the focus is on two lists which are chosen by the user.

In this more relaxed set-up, the target parameters are only partially identified instead of point identified, unlike the set-up of Das et al. (2021). Estimation under only partial identification has gained interest in the last few years (Imbens and Manski, 2004). The standard approach in this case is to estimate a range for the parameter of interest instead of a point estimate. Imbens and Manski (2004) provides a method to calculate confidence intervals for the parameter when the distributions of the estimated upper and lower bounds of the parameter are available. We apply this approach in the context of this paper and further present the finite sample error in the coverage guarantees of the proposed confidence interval.

4.1.1 Overview of the paper

In this paper, we discuss estimation of the total population size when the two lists deviate from conditional independence assumption in the presence of covariates. Section 4.2 describes the set-up and data structure. In section 4.3, we present the general set-up of partial identification as in Imbens and Manski (2004) and present the finite sample coverage error in the estimated confidence interval. In section 4.4, we discuss modelling the dependence using conditional risk ratio of the two lists and present the upper and lower bounds of the capture probability along with variance estimators. Following this, in section 4.5, we present estimators for the total population size and present the finite sample coverage error of the proposed confidence interval. Section 4.6 discusses the performance of the proposed method in a simulated set-up and presents interval estimates of the total number of victims in the Peru Internal Armed Conflict of 1980-2000 for various levels of relaxation of the conditional independence assumption.

4.2 Preliminaries

4.2.1 Set-up

The population size estimation in this paper is in a capture-recapture set-up. We use the set-up from Das et al. (2021). Consider a finite population of n individuals. Suppose the samples consists of data from K lists. An individual is observed if he is captured by at least one of the K lists. Denote the number of the observed individuals by N. For individual i, $i \in \{1, \ldots, n\}$, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})^T$ denotes the indicator vector for the capture-history. $Y_{ik} \in \{1, 0\}$ is the indicator of whether individual i is captured in list k or not, $k \in \{1, \ldots, K\}$. We also consider covariates $\mathbf{X}_i \in \mathbb{R}^d$ for each individual. We assume that every individual behaves independently of the other individuals. Hence, the vector $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i)$ are independent and identically distributed according to some distribution $\mathbb{P}$.

If all the individuals in the population are observed, then $n = N$ and we do not need to estimate anything. However, in practice there are a substantial number of individuals not captured by any of the lists, i.e. $\mathbf{Y} = \mathbf{0}$. We however have access to only N individuals, where $N = \sum_{i=1}^{n} \mathbb{1}(\mathbf{Y}_i \neq \mathbf{0})$. The observed data size, N follows $Binomial(n, \psi)$, where $\psi = \mathbb{P}(\mathbf{Y} \neq \mathbf{0})$. We are interested in the estimation of n which is equivalent to the estimation of ψ. By the property of binomial distribution, any estimator of ψ, say $\widehat{\psi}$ gives an estimator of n by $N/\widehat{\psi}$. the estimation of ψ and n are equivalent.

By structure, capture-recapture is a missing data problem where the observed data is possibly a biased-sample from the population. Hence, we cannot estimate $\mathbb{P}$ directly from the observed data. The observed data, however, follows a conditional distribution $\mathbb{Q}$, i.e. $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i) \sim \mathbb{Q}$, for $i \in \{1, \ldots, N\}$.

$$\begin{aligned}\mathbb{Q}(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}) &\equiv \mathbb{P}(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x} | \mathbf{Y} \neq \mathbf{0}) \\ &= \psi^{-1} \mathbb{P}(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}) \mathbb{1}(\mathbf{y} \neq \mathbf{0}).\end{aligned}$$

Das et al. (2021) used the assumption of conditional independence between two lists to ensure identification of ψ and n. In this paper, we relax this assumption and develop confidence interval estimates under partial identification.

4.3 Estimation of target parameter using the estimated bounds

All the estimators presented in Das et al. (2021) are valid only when there are two lists that are independent conditional on the covariates. In general, for real data this identification assumption may not hold true. For example, the corresponding list pair may not be conditionally independent. Unless we know the extent of dependence between the two lists, there is no point identification. We present two approaches for estimation under this violation in section 4.4.

In case of partial identification, usually one has a range of estimates instead of a point estimate. The standard approach is to use the range of estimates to obtain a confidence interval for the parameter of interest. Imbens and Manski (2004) has presented a general formula for the calculation of the confidence interval of the target parameter when we know the estimators for the upper and lower bounds of the parameter; and further they are asymptotically normal. In this section we briefly discuss the approach of Imbens and Manski (2004) and present the finite sample coverage error of the estimated confidence interval.

Consider a general partial identification problem with target parameter ψ and the population upper and lower bounds for ψ are ψ_u and ψ_l respectively. Also, let the estimated range be $(\widehat{\psi}_l, \widehat{\psi}_u)$. Suppose the following is the asymptotic joint distribution of the estimated maximum and the estimated minimum. When there are N many observations in the data, we have the following.

$$\sqrt{N}\begin{pmatrix}\widehat{\psi}_l - \psi_l \\ \widehat{\psi}_u - \psi_u\end{pmatrix} \longrightarrow N\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_l^2 & cov \\ cov & \sigma_u^2\end{pmatrix}\right\},$$

where σ_l and σ_u are the corresponding standard deviations. The covariance is not necessary for this set-up.

According to Imbens and Manski (2004), one can construct a $(1-\omega)\%$ confidence interval for ψ as follows

$$\bar{CI}^{\psi}_{1-\omega} = \left[\widehat{\psi}_l - \bar{C}_N\frac{\widehat{\sigma}_l}{\sqrt{N}}, \widehat{\psi}_u + \bar{C}_N\frac{\widehat{\sigma}_u}{\sqrt{N}}\right],$$

where $\bar{C}_N$ satisfies $\Phi\left(\bar{C}_N + \sqrt{N}\frac{\widehat{\psi}_u - \widehat{\psi}_l}{\max(\widehat{\sigma}_l, \widehat{\sigma}_u)}\right) - \Phi\left(-\bar{C}_N\right) = 1-\omega$. Imbens and Manski (2004) has proved that this estimated confidence interval contains the true target parameter with probability at least $1-\omega$ asymptotically.

In general, the estimated parameters need not be asymptotically normal and may not be unbiased. Hence, we consider a general case with bias and approximate normality and present the finite sample error in the coverage probability of the estimated interval for a given sample size N.

Theorem 4.1. *Suppose the estimated lower and upper bounds $\widehat{\psi}_l$ and $\widehat{\psi}_u$ satisfy the following.*

1. $\mathbb{E}(\widehat{\psi}_l) = \psi_l + \widehat{R}_{2,l}$
2. $\mathbb{E}(\widehat{\psi}_u) = \psi_u + \widehat{R}_{2,u}$
3. $var(\widehat{\psi}_l|\mathbf{Z}^n) = \widetilde{\sigma}_l^2/N$, $\widehat{var}(\widehat{\psi}_l) = \widehat{\sigma}_l^2/N$
4. $var(\widehat{\psi}_u|\mathbf{Z}^n) = \widetilde{\sigma}_u^2/N$, $\widehat{var}(\widehat{\psi}_u) = \widehat{\sigma}_u^2/N$.

$\widehat{R}_{2,l}$ is the bias, $\widetilde{\sigma}^2$ is the population variance given the training sample, and $\widehat{\sigma}^2$ is the estimated variance for the corresponding bound. Further, assume that $\widehat{\psi}_u$ and $\widehat{\psi}_l$ are sample averages of i.i.d terms and uniformly continuous. Let ρ_l and ρ_u be the absolute central third moment of the i.i.d terms. Moreover, $\mathbb{E}|\widehat{\sigma} - \widetilde{\sigma}| \lesssim n^{-1/2}$. For simplicity, define $\Delta = \psi_u - \psi_l$ and $\widehat{\Delta} = \widehat{\psi}_u - \widehat{\psi}_l$. Suppose the following assumption holds.

Assumption 3. *For a given $\epsilon > 0$ and a constant c, there exists N_0 and $\upsilon > 0$ such that for all $N > N_0$*

$$\mathbb{P}\left(N^{\upsilon}|\widehat{\Delta} - \Delta| > c\right) < \epsilon.$$

We use $\epsilon < 1/\sqrt{n}$ for a given n. Further, let $\underline{\sigma} \leq \widehat{\sigma}_l, \widetilde{\sigma}_l, \widehat{\sigma}_u, \widetilde{\sigma}_u \leq \bar{\sigma}$ for some finite real numbers $\underline{\sigma}$ and $\bar{\sigma}$. Then the finite sample lower bound on the empirical coverage is presented below.

$$
\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N \widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \bar{C}_N \widehat{\sigma}_u/\sqrt{N}\right) \\
\leq\; & \frac{1}{\sqrt{2\pi}\underline{\sigma}_l} \mathbb{E}\left(\sqrt{N}|\widehat{R}_{2,l}| + \sqrt{N}|\widehat{R}_{2,u}|\right) + \mathbb{E}\left\{\frac{C}{\sqrt{N}}\left(\frac{\rho_l}{\widetilde{\sigma}_l^3} \vee \frac{\rho_u}{\widetilde{\sigma}_u^3}\right)\right\} \\
&+\; \mathbb{E}\left(\frac{c_N}{\sqrt{2\pi}\underline{\sigma}_l}|\widehat{\sigma}_u - \widetilde{\sigma}_u|\right) + \mathbb{E}\left(\frac{1+c_N}{\sqrt{2\pi}\underline{\sigma}_l}|\widehat{\sigma}_l - \widetilde{\sigma}_l|\right) \\
&+\; \mathbb{1}(\Delta \neq 0)\mathbb{E}\left\{\frac{1}{\sqrt{N}}(1+4\theta) + \frac{2c\bar{\sigma}^2}{\alpha N^{\frac{1}{2}+\upsilon}\,\Delta^3} + \left|\frac{\bar{\sigma}_l}{\underline{\sigma}_l} - 1\right| \frac{\bar{\sigma}_l}{\sqrt{N}\Delta}\right\} \\
&+\; \mathbb{1}(\Delta \neq 0)\mathbb{E}\left(\frac{3}{\alpha\widehat{\Delta}\underline{\sigma}_l}|\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|\right).
\end{aligned}
$$

where θ is the maximum value of the density of $\sqrt{N}(\widehat{\psi}_u - \psi_u)/\widehat{\sigma}_u$ and $\sqrt{N}(\widehat{\psi}_l - \psi_u)/\widehat{\sigma}_l$. By Berry-Esseen, $\theta \approx 1/\sqrt{2\pi}$. Moreover, $\bar{C}_N$, $c_N \leq z_{1-\alpha/2}$.

The above theorem says that the estimated confidence interval contains the true parameter ψ_u with probability at least $1-\alpha$ minus some additional error term which is bounded above. One can obtain a similar result for the lower bound ψ_l. Thus, since the estimated confidence interval contains both ψ_u and ψ_l with some probability, it must also contain ψ with probability at least $1-\alpha$ with some bounded error term. This is a general result that is applicable to any partial identification problem that follows the basic set-up presented in Imbens and Manski (2004) and in this paper. A more precise bound is presented in the proof.

The assumption used in the above theorem says that the estimated difference $\widehat{\Delta} = \widehat{\psi}_u - \widehat{\psi}_l$ is converging to the true difference Δ roughly at rate $o(N^{-\upsilon})$ for some $\upsilon > 0$. This is a mild assumption that is easily satisfied in most estimation scenarios.

The primary result in Imbens and Manski (2004) shows that this interval contains the true parameter with probability at least $1-\alpha$ for sufficiently large sample size. The theorem above quantifies that coverage error for any sample size N. This error further decreases as N (also, equivalently n) increases under some weak conditions on the bias terms $\widehat{R}_{2,l}$ and $\widehat{R}_{2,u}$. We discuss these conditions in a later section in the context of the proposed estimators in this paper. Below, we present the large sample error bound in the following corollary.

Corollary 4.1.1. *Assume $\underline{\sigma} \gtrsim 1$, $\mathbb{E}|\widehat{R}_{2,l}| \vee \mathbb{E}|\widehat{R}_{2,u}| \lesssim N^{-2\beta}$ and $\alpha > 0$. Then the coverage error for the proposed $(1-\alpha)$ confidence interval is upper bounded by*

$$
(1-\alpha) - \mathbb{P}\left(\widehat{CI} \ni \psi^{-1}\right) \;\lesssim\; n^{(1-4\beta)/2} + \frac{1}{\sqrt{n}}.
$$

Therefore, if $\beta > 1/4$ there exists some sample size N_ϵ at which the coverage error is never more than ϵ, for any $N > N_\epsilon$.

4.4 Estimation under dependence

In this section, we consider relaxations of the conditionally independent assumption. For simplicity, we will focus on only two lists. Without loss of generality, let the lists be list 1 and 2. As mentioned before also stated in Das et al. (2021), for estimation of the total population size or the capture probability, we need some lack of dependence assumption between the lists under consideration. Hence, to proceed with the estimation, we assume that lists 1 and 2 are not completely dependent. This dependence can be measured by either risk ratio or odds ratio. Under independence, both risk and odds ratios are 1. However, for real data, this may not hold true. In this paper, we will focus on the risk ratio. One can derive equivalent results for the odds ratio. When the two lists are not conditionally independent, the risk ratio can deviate from 1. Instead of assuming a fixed value of dependence, i.e. the risk ratio, we use a weak assumption that the risk ratio lies in a neighbourhood of 1. This relaxation does not guarantee point identification. Hence, instead of estimating the target parameter, we estimate the upper and lower bounds of the parameter.

The mild identification assumption presented below assumes that the risk ratio lies in a neighbourhood of 1. This is a relaxation of the conditional independence assumption used in Das et al. (2021) and is more difficult.

Assumption 4. *The risk ratio between lists 1 and 2 is bounded as follows for some finite $\omega > 1$.*

$$\frac{1}{\omega} \leq \frac{\mathbb{P}(Y_1 = 1 \mid Y_2 = 1, \mathbf{X} = \mathbf{x})}{\mathbb{P}(Y_1 = 1 \mid Y_2 = 0, \mathbf{X} = \mathbf{x})} \leq \omega, \, \forall \, \mathbf{x}.$$

The ratio in the above assumption is the risk ratio between lists 1 and 2 conditional on the covariate. When $\omega = 1$, the risk ratio is 1 for all $\mathbf{x}$'s. Thus, $\omega = 1$ implies conditional independence between lists 1 and 2 i.e., $Y_1 \perp\!\!\!\perp Y_2 \mid \mathbf{X}$. As ω increases, the assumption becomes weaker. For meaningful implementation of this assumption, one has to choose ω not too close to ∞.

Under this assumption, we derive a reasonable bound for the inverse capture probability ψ^{-1}. since the total population size is a linear function of the inverse capture probability, we present all the results for the inverse capture probability instead of the capture probability. First, we begin by expressing the inverse capture probability as a function of the observed data and the risk ratio. The following proposition presents the result.

Proposition 2. *For a risk ratio function* $\delta(\mathbf{x}) = \frac{\mathbb{P}(Y_1=1|Y_2=1,\mathbf{X}=\mathbf{x})}{\mathbb{P}(Y_1=1|Y_2=0,\mathbf{X}=\mathbf{x})}$*, the conditional capture probability is*

$$\gamma(\mathbf{x}) = \frac{q_{12}(\mathbf{x})}{[\{q_1(\mathbf{x}) - q_{12}(\mathbf{x})\}\delta(\mathbf{x}) + q_{12}(\mathbf{x})]q_2(\mathbf{x})}.$$

The capture probability ψ *associated with the above* γ *is*

$$\psi_\delta^{-1} = \int \frac{1}{\gamma(\mathbf{x})} d\mathbb{Q}(\mathbf{x}) = \int \Big[\delta(\mathbf{x})\{q_1(\mathbf{x}) - q_{12}(\mathbf{x})\} + q_{12}(\mathbf{x})\Big] \frac{q_2(\mathbf{x})}{q_{12}(\mathbf{x})} d\mathbb{Q}(\mathbf{x}).$$

The above expression shows that the inverse capture probability increases with the risk ratio $\delta(\mathbf{x})$. Note that $q_1(\mathbf{x})$, $q_2(\mathbf{x})$ and $q_{12}(\mathbf{x})$ can be directly estimated from the observed data, $\delta(\mathbf{x})$ cannot. Hence, one cannot obtain a point estimate of the capture probability from the observed data without using additional information about the risk ratio $\delta(\mathbf{x})$. Next, we show how using assumption 4, one can obtain bounds on ψ, or equivalently ψ^{-1}.

Remark 24. *When the two lists are conditionally independent, i.e.,* $\delta(\mathbf{x}) = 1 \ \forall \ \mathbf{x}$*, then the conditional inverse capture probability is* $\frac{q_1(\mathbf{x})q_2(\mathbf{x})}{q_{12}(\mathbf{x})}$*. In the relaxed setup, depending on how list 2 affects list 1,* $\delta(\mathbf{x})$ *scales the proportion that is observed by only list 1 and not by list 2. For example, if we believe that* $\delta(\mathbf{x}) < 1$*, i.e., being on list 2 decreases the chances of being on list 1, then* $q_1(\mathbf{x}) - q_{12}(\mathbf{x})$ *is larger relative to* $q_{12}(\mathbf{x})$*.*

Assumption 4 can only ensure partial identification of ψ, and hence ψ^{-1}, since we do not assume point values for the risk ratios. Let ψ_l^{-1} and ψ_u^{-1} denote the lowest and the highest values ψ^{-1} can attain under assumption 4. Under this assumption, the true parameter ψ^{-1} should lie between ψ_l^{-1} and ψ_u^{-1} because of the monotone nature of the identifiable expression. We present these bounds in the following theorem.

Lemma 4.1.1. *Under assumption 4, the lower and upper bounds on* ψ^{-1} *are as follows.*

$$\begin{aligned}
\psi_l^{-1} &= \int \left\{ \frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1 \right\} \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1 \right\} d\mathbb{Q}(\mathbf{x}) + 1 \\
\psi_u^{-1} &= \int \left\{ \frac{1}{\gamma_\omega(\mathbf{x})} - 1 \right\} \mathbb{1}\left\{ \gamma_\omega(\mathbf{x}) \leq 1 \right\} d\mathbb{Q}(\mathbf{x}) + 1,
\end{aligned}$$

where $\gamma_\omega^{-1}(\mathbf{x}) = [\omega\{q_1(\mathbf{x}) - q_{12}(\mathbf{x})\} + q_{12}(\mathbf{x})] \frac{q_2(\mathbf{x})}{q_{12}(\mathbf{x})}$ *is the inverse conditional capture probability when the risk ratio is* ω*. When* $\omega = 1$*, i.e., the two lists are conditionally independent, then*

$$\psi_1^{-1} = \psi_u^{-1} = \int \frac{q_1(\mathbf{x})q_2(\mathbf{x})}{q_{12}(\mathbf{x})} d\mathbb{Q}(\mathbf{x}) = \int \frac{1}{\gamma_1(\mathbf{x})} d\mathbb{Q}(\mathbf{x}).$$

The final expression of the bounds are obtained by adjusting the risk ratio $\delta(\mathbf{x})$ such that $\gamma(\mathbf{x}) \leq 1$ for all $\mathbf{x}$, followed by rearrangement of the terms. The bounds above are sharper in the sense that, we are selecting the maximum (or minimum) of the inverse capture probabilities for each $\mathbf{x}$, i.e., for each individual. Further, we ensure that the inverse capture probabilities are greater than or equal to 1. As a consequence, the integrands are not smooth functions of the $\mathbf{x}$'s. However, if γ is sufficiently smooth, then ψ_l^{-1} and ψ_u^{-1} are smooth functions of ω.

Moreover, it is easy to see that as ω increases or, equivalently, $1/\omega$ gets closer to 0, ψ_l^{-1} decreases and ψ_u^{-1} increases. Thus, the bounds grow further apart. Thus, these bounds are also monotone in ω.

4.4.1 Efficiency bound

In this section, we derive the nonparametric efficiency bounds for estimation using i.i.d. samples from the observed data distribution $\mathbb{Q}$. This bound sets the benchmark for the best possible variance an estimator can achieve. To evaluate the efficiency bounds, we begin by deriving the efficient influence functions for ψ_l^{-1} and ψ_u^{-1} (Bickel et al., 1993).

The efficient influence function of a parameter quantifies the change in the parameter when one introduces perturbations in the input distribution. The efficient influence function has many important properties. Its variance gives the efficiency bound, which sets the lowest possible variance an estimator can achieve. And once, we obtain an estimator that achieves this bound, no further improvement scan be made in terms of the bound. The efficient influence function is used to construct an estimator that achieves this bound under some regularity conditions (Bickel et al., 1993; van der Vaart, 2002a; van der Laan and Robins, 2003; Tsiatis, 2006; Kennedy, 2016).

Existence of the efficient influence function, requires that the parameter is sufficiently smooth. However, identifiable expressions for ψ_l^{-1} and ψ_u^{-1} contain indicator terms involving the conditional capture probability. Hence, we use the following margin assumption to ensure that $\gamma_\omega(\mathbf{x})$ is sufficiently smooth around 1 for any positive value of ω.

Assumption 5. *(Margin) There exists a constant* $\nu > 0$*, such that for all* $t > 0$*,* $\mathbb{Q}(|\gamma_\omega - 1| \leq t) \lesssim t^\nu$*.*

In the following theorem, we evaluate the efficient influence functions for ψ_l^{-1} and ψ_u^{-1} and further, we have shown in the proof in the appendix that they exist under the above margin condition.

Theorem 4.2. *Given that assumption 5 holds and $\nu \geq 1$. Then the efficient influence functions for ψ_l^{-1} and ψ_u^{-1} respectively are*

$$\begin{aligned}\phi_l(\mathbf{Z};\mathbb{Q}) &= \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\left(\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})}\left[\frac{(Y_1 - Y_1Y_2)/\omega + Y_1Y_2}{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}/\omega + q_{12}(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right] - 1\right) \\ &\quad + 1 - \psi_l^{-1}, \text{ and} \\ \phi_u(\mathbf{Z};\mathbb{Q}) &= \mathbb{1}\left\{\gamma_{\omega}(\mathbf{X}) \leq 1\right\}\left(\frac{1}{\gamma_{\omega}(\mathbf{X})}\left[\frac{(Y_1 - Y_1Y_2)\omega + Y_1Y_2}{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}\omega + q_{12}(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right] - 1\right) \\ &\quad + 1 - \psi_u^{-1}.\end{aligned}$$

The above functions are efficient influence functions of ψ_l^{-1} and ψ_u^{-1} respectively when $\nu \geq 1$ in the margin condition in assumption 5. The efficient influence functions above have zero means. Hence, they can be used to derive alternate expressions for ψ_l^{-1} and ψ_u^{-1}. The estimated efficient influence functions are obtained by replacing the q-probabilities with their respective estimates.

Remark 25. *For simplicity, we will use ϕ_l and ϕ_u to denote $\phi_l(\mathbf{Z};\mathbb{Q})$ and $\phi_u(\mathbf{Z};\mathbb{Q})$ respectively. And, $\widehat{\phi}_l$ and $\widehat{\phi}_u$ denote the respective estimates $\phi_l(\mathbf{Z};\widehat{\mathbb{Q}})$ and $\phi_u(\mathbf{Z};\widehat{\mathbb{Q}})$, i.e., when one used the distribution $\widehat{\mathbb{Q}}$ when the true distribution is $\mathbb{Q}$.*

The variance of the efficient influence sets the benchmark against which one can check the performance of an estimator of ψ_l^{-1} and ψ_u^{-1}. In the following corollary we present the variance expression of the efficient influence function associated with ψ_l^{-1}. The equivalent expression for ψ_u^{-1} is obtained by replacing ω with $1/\omega$.

Corollary 4.2.1. *The variance of ϕ_l is*

$$\begin{aligned}var(\phi_l) &= \mathbb{E}\Bigg(\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\Bigg[\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})}\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\}\left\{\frac{1}{\omega q_{12}(\mathbf{X})} - 1\right\} + \frac{q_0(\mathbf{X})}{\omega\gamma_{\frac{1}{\omega}}(\mathbf{X})q_{12}(\mathbf{X})} \\ &\quad + \left(1 - \frac{1}{\omega}\right)\frac{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}^2 q_2(\mathbf{X})^2}{\omega^2 q_{12}(\mathbf{X})^3}\Bigg]\Bigg) + var\left[\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\}\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right], \\ var(\phi_u) &= \mathbb{E}\Bigg(\mathbb{1}\left\{\gamma_{\omega}(\mathbf{X}) \leq 1\right\}\Bigg[\frac{1}{\gamma_{\omega}(\mathbf{X})}\left\{\frac{1}{\gamma_{\omega}(\mathbf{X})} - 1\right\}\left\{\frac{\omega}{q_{12}(\mathbf{X})} - 1\right\} + \frac{\omega q_0(\mathbf{X})}{\gamma_{\omega}(\mathbf{X})q_{12}(\mathbf{X})} \\ &\quad - \omega^2\ (\omega - 1)\frac{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}^2 q_2(\mathbf{X})^2}{q_{12}(\mathbf{X})^3}\Bigg]\Bigg) + var\left[\left\{\frac{1}{\gamma_{\omega}(\mathbf{X})} - 1\right\}\mathbb{1}\left\{\gamma_{\omega}(\mathbf{X}) \leq 1\right\}\right],\end{aligned}$$

where $q_0(\mathbf{x}) = 1 - q_1(\mathbf{x}) - q_2(\mathbf{x}) + q_{12}(\mathbf{x})$ is the probability of being observed by neither list 1 nor list 2.

The variance above is influenced by five main factors:

1. the deviation from conditional independence ω,
2. the conditional capture probability $\gamma_{\frac{1}{\omega}}(\mathbf{x})$,

3. the probability of appearing only on list 1, $q_1(\mathbf{x}) - q_{12}(\mathbf{x})$,
4. the probability of appearing on both list 1 and 2,
5. the heterogeneity in the conditional capture probabilities $var\left[\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right]$.

4.4.2 Estimation

In lemma 4.1.1, we derived the identifiable expressions for ψ_l^{-1} and ψ_u^{-1}. In this section, we will present estimators for the bounds ψ_l^{-1} and ψ_u^{-1}. The most straight-forward estimators are the plug-in estimators, which are obtained by the sample analogues of the identifiable expressions in lemma 4.1.1. But, plug-in estimators typically have first-order bias, and moreover, in a nonparametric set-up, they have slow convergence rates and no well-defined variance formula, and hence, often do not have a known asymptotic distribution. Hence, we turn to efficiency theory to overcome these shortcomings.

The plug-in estimator for the bounds on ψ^{-1} are obtained by substituting the nuisance functions with their estimates in the expressions in lemma 4.1.1 as follows.

$$\begin{aligned}
\widehat{\psi}_{l,pi}^{-1} &= \mathbb{Q}_N \left[\left\{\frac{1}{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\} \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right] + 1. \\
\widehat{\psi}_{u,pi}^{-1} &= \mathbb{Q}_N \left[\left\{\frac{1}{\widehat{\gamma}_{\omega}(\mathbf{X})} - 1\right\} \mathbb{1}\left\{\widehat{\gamma}_{\omega}(\mathbf{X}) \leq 1\right\}\right] + 1.
\end{aligned}$$

As discussed above, these plug-in estimators have some disadvantages, that need to be addressed. These can be solved by using the efficient influence functions presented in theorem 4.2. The efficient influence functions have quite a few desirable properties: (i) They can be used to construct alternate bias-corrected estimators for the parameters ψ_l^{-1} and ψ_u^{-1}, (ii) The second-order remainder term can be used to quantify the error in estimation and study the robustness of the estimator, (iii) The variance of the efficient influence functions can be used to find the lowest possible variance an estimator can achieve.

Using the efficient influence function above and efficiency theory from Tsiatis (2006); Kennedy (2016); van der Vaart (2002b); Bickel et al. (1993), we can obtain a proposed estimators of ψ_l^{-1} and ψ_u^{-1}.

$$\begin{aligned}
\widehat{\psi}_{l,proposed}^{-1} &= \widehat{\psi}_{l,pi}^{-1} + \mathbb{Q}_N \widehat{\phi}_l, \\
\widehat{\psi}_{u,proposed}^{-1} &= \widehat{\psi}_{u,pi}^{-1} + \mathbb{Q}_N \widehat{\phi}_u,
\end{aligned}$$

where $\widehat{\phi}$ are the corresponding estimated efficient influence functions. We obtain the above estimators by using the property that efficient influence functions have mean zero. $\mathbb{Q}_N \widehat{\phi}_l$ is the estimated first-order bias in the plug-in estimator. Hence, the proposed estimators also do not have first-order bias.

Note that as a consequence of using the efficient influence function, the error in the estimation when using the proposed estimators can be approximated as a sample average of i.i.d. random variables. Approximation with a sample average makes it easier to evaluate the variance. In the next result, we derive the bound on this approximation error.

Theorem 4.3. *For any sample size N and error tolerance $\eta > 0$, we have*

$$\left|\widehat{\psi}_{l,proposed}^{-1} - \psi_l^{-1} - \mathbb{Q}_N \phi_l\right| \leq \eta,$$

with probability at least

$$1 - \frac{1}{\eta^2}\mathbb{E}\left[\widehat{R}_{2,l}^2 + \frac{1}{N}\left\|\widehat{\phi}_l - \phi_l\right\|\right],$$

where $\widehat{R}_{l,2}$ is a second-order error term given by

$$\begin{aligned}
\widehat{R}_{2,l} = &\int \mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{1}{\omega \bar{q}_{12}(\mathbf{x})}\Bigg[\{q_1(\mathbf{x}) - \bar{q}_1(\mathbf{x})\}\{\bar{q}_2(\mathbf{x}) - q_2(\mathbf{x})\} \\
&+ \{q_{12}(\mathbf{x}) - \bar{q}_{12}(\mathbf{x})\}\left\{\frac{q_2(\mathbf{x})q_1(\mathbf{x})}{q_{12}(\mathbf{x})} - \frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})}\right\}\Bigg] d\mathbb{Q}(\mathbf{x}) \\
&+ \int \left[\mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} - \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\}\right]\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1\right\} d\mathbb{Q}(\mathbf{x}) \\
\leq &\left(\frac{1}{\omega\epsilon}\right)\|\widehat{q}_1 - q_1\|\|\widehat{q}_2 - q_2\| + \left(\frac{1}{\omega\epsilon}\right)\|\widehat{q}_{12} - q_{12}\|\left\|\frac{q_2 q_1}{q_{12}} - \frac{\widehat{q}_2\widehat{q}_1}{\widehat{q}_{12}}\right\| \\
&+ \frac{1}{\epsilon}\left\|\bar{\gamma}_{\frac{1}{\omega}} - \gamma_{\frac{1}{\omega}}\right\|_\infty^{1+\nu}.
\end{aligned}$$

where the latter bound holds as long as $(q_{12} \wedge \widehat{q}_{12}) \geq \epsilon$ and assumption 5 holds.

The above theorem shows that the error in estimation and hence, the proposed estimator can be approximated by a sample average of a function of the efficient influence function. In assumption 5, when $\nu \geq 1$, $\widehat{R}_{2,l}$ is second order. This approximation falls within a tolerance η with a probability that depends on the second order error term and the standard deviation of $\widehat{\phi}_l$. Further, the probability increases as the sample size increases or the remainder term decreases.

This second order error term summarizes the estimation error in ψ_l. Similarly, one can conclude for $\widehat{R}_{2,u}$. These error terms also govern the coverage error in the estimated confidence interval presented in section 4.3. If the risk ratio $\zeta(\mathbf{x})$ is s-Holder-continuous, then $\|\widehat{\gamma}_{\frac{1}{\omega}} - \gamma\|_\infty = O(n^{-\frac{s}{1-s}})$, where the nuisance parameters are estimated with error rate $n^{-\beta}$. Thus, $\mathbb{E}|\widehat{R}_{2,l}| \lesssim n^{-2\beta} + n^{-\frac{(1+\nu)s}{1-s}}$. The coverage error in theorem 4.1 decreases with N and equivalently n if $\beta > 1/4$ and $(1+\nu)s/(1-s) > 1/2$ or $\nu > (1-3s)/(2s)$. For second-order error and for ϕ_l to be valid efficient influence functions, we need that ν is at least 1.

As a consequence of the above theorem, the error in estimation is now approximated by a sample average of the efficient influence function ϕ_l. Thus, one can now evaluate the variances for the estimators via the variances of the efficient influence functions presented in corollary 4.2.1. To formally write, the variance of $\widehat{\psi}_l^{-1}$ is approximated as follows.

$$var(\widehat{\psi}_l^{-1}) \approx \frac{1}{N} var\left(\widehat{\phi}_l\right).$$

Since, $\widehat{\psi}_l^{-1}$ and $\widehat{\psi}_u^{-1}$ are approximately sample averages, one can show that they are approximately normal. For more details, we refer to Das et al. (2021). This allows us to implement the Imbens and Manski (2004) confidence interval formula to construct a confidence interval for the target parameter ψ (equivalently ψ^{-1}). We have presented the general result along with the coverage guarantees in section 4.3.

In this section, we considered a weak identifying assumption that the risk ratio is bounded; which further guarantees only partial identification. Under this assumption, we presented the identifying expressions for the lower and upper bounds of the capture probability ψ (equivalently ψ^{-1}) and derived the efficiency bounds. Further, we presented the doubly robust estimators, i.e. estimators with second-order error terms for the bounds that achieve the efficiency bound under a margin condition to ensure smoothness. Lastly, we discussed how one can construct a confidence interval for ψ (equivalently ψ^{-1}) using Imbens and Manski (2004) formula. In the following section, we discuss the inferences for the total population size.

4.5 Confidence interval for the total population size

In this section, we lay out the steps to obtain a confidence interval for the total population size n using the results from the previous sections. Under assumption 4, n is not identifiable from the observed data. Hence, we define the lower and upper bounds for n and present their estimators and variances, which allow us to apply the Imbens and Manski (2004) formula. Further, we present the coverage guarantee for the proposed estimated confidence interval for n.

Under assumption 4, we defined the identifiable range for the total population size by the interval (n_l, n_u). Suppose n_l and n_u satisfy the following for consistency.

$$n_l = n\psi\psi_l^{-1}, \qquad n_u = n\psi\psi_u^{-1}.$$

Using the above identifiable expressions, one can use the derived results for ψ to infer about the bounds on n. The results presented in this section hold for all estimators of ψ_u and ψ_l that satisfy a very mild condition presented in the theorem.

Given any estimators for the bounds of ψ^{-1}, one can obtain the respective estimators for n_l and n_u by

$$\widehat{n}_l = N\widehat{\psi}_l^{-1}, \qquad \widehat{n}_u = N\widehat{\psi}_u^{-1}. \tag{4.1}$$

The estimators of n_l and n_u are products of two random quantities: (i) the observed data size N, and (ii) the estimated bound on ψ^{-1}. Hence, to derive the variance, and hence, the coverage error of the confidence interval, it requires some non-trivial extension of the results in the previous sections. In the following theorem, we present the variances for the estimated bounds of n.

Theorem 4.4. *Let $\widehat{\psi}_l^{-1}$ and $\widehat{\psi}_u^{-1}$ be generic estimators for ψ_l^{-1} and ψ_u^{-1}, respectively, that satisfy*

$$\widehat{\psi}_l^{-1} - \psi_l^{-1} = \mathbb{Q}_N(\widehat{\varphi}_l) - \int \varphi_l(\mathbf{z}) d\mathbb{Q}(\mathbf{z}) + \widehat{R}_{2,l}, \tag{4.2}$$

where φ_l is a generic influence function and $\widehat{R}_{2,l}$ is the corresponding second order error term.$\widehat{\psi}_u^{-1}$ has a similar expression. Let $\widetilde{\varsigma}_l^2 = var(\widehat{\varphi}_l|\mathbf{Z}^n)$ and $\widetilde{\varsigma}_u^2 = var(\widehat{\varphi}_u|\mathbf{Z}^n)$ be the corresponding efficiency bounds. Then the variance of $\widehat{n}_l$ and $\widehat{n}_u$ are

$$\begin{aligned} var(\widehat{n}_l) &= n\psi \left\{ n\psi \ var(\widehat{R}_{2,l}) + \frac{1-\psi}{\psi_l^2} \mathbb{E}(\psi_l \widehat{R}_{2,l} + 1)^2 + \mathbb{E}(\widetilde{\varsigma}_l^2) \right\} \\ var(\widehat{n}_u) &= n\psi \left\{ n\psi \ var(\widehat{R}_{2,u}) + \frac{1-\psi}{\psi_u^2} \mathbb{E}(\psi_u \widehat{R}_{2,u} + 1)^2 + \mathbb{E}(\widetilde{\varsigma}_u^2) \right\}. \end{aligned}$$

The above theorem says that the variances of the estimated lower and upper bounds of n (i) increase with the variance of the influence functions, which also summarize the variance of the estimators of ψ_u^{-1} and ψ_l^{-1}, (ii) increase with the remainder term and its variance, (iii) depend on the capture probability ψ but the trend is not clearly monotone, and (iv) decrease with ψ_u and ψ_l. Further, it is important to note that the variances are of order n if the remainder terms and their variances are sufficiently small. The conditions for the later are discussed in the previous section following theorem 4.3. These variance formulas hold for any general estimators for the bounds of ψ^{-1}that satisfy the mild condition 4.2 presented in the above theorem.

Next, we will present the estimators of these variances which will be used in the construction of confidence interval for n. The above variance formulas contain the true capture probability ψ and the total population size n. One can approximate $n\psi$ by the number of observations N using the Binomial assumption discussed in section 4.2. As for the $1-\psi$ in the second term, we substitute it by $1-\widehat{\psi}_u$ to ensure maximum coverage by the confidence interval. By our definition, $\psi_u \leq \psi_l$, since they are respectively the upper and the lower bounds of the inverse capture probability. We estimate $\widetilde{\varsigma}_l^2$ and $\widetilde{\varsigma}_u^2$ as follows.

$$\widehat{var}(\widehat{n}_l) = N\widehat{\tau}_l^2 = N\left(\widehat{\varsigma}_l^2 + \frac{1-\widehat{\psi}_u}{\widehat{\psi}_l^2}\right),$$

$$\text{and } \widehat{var}(\widehat{n}_u) = N\widehat{\tau}_u^2 = N\left(\widehat{\varsigma}_u^2 + \frac{1-\widehat{\psi}_u}{\widehat{\psi}_u^2}\right),$$

where $\widehat{\tau}_l$ denotes $\widehat{\varsigma}_l^2 + \frac{1-\widehat{\psi}_u}{\widehat{\psi}_l^2}$. Similarly, we have $\widehat{\tau}_u$. Note that these variance estimates are positively biased. We further define the true variances for $\widehat{n}_l$ and $\widehat{n}_u$ conditional on the training sample as follows.

$$\begin{aligned} var(\widehat{n}_l|\mathbf{Z}^n) &= n\psi\widetilde{\tau}_l^2 = n\psi\widehat{\varsigma}_l^2 + n\frac{\psi(1-\psi)}{\psi_l^2}(\psi_l\widehat{R}_{2,l}+1)^2, \\ \text{and } var(\widehat{n}_u|\mathbf{Z}^n) &= n\psi\widetilde{\tau}_u^2 = n\psi\widehat{\varsigma}_u^2 + n\frac{\psi(1-\psi)}{\psi_u^2}(\psi_u\widehat{R}_{2,u}+1)^2. \end{aligned}$$

The above expressions follow from the fact that $\widehat{R}_{2,l}$ and $\widehat{R}_{2,u}$ are constants when conditioning on the training sample $\mathbf{Z}^n$. We use these expressions, to calculate the finite sample coverage guarantees of the proposed confidence interval.

To construct the $(1-\alpha)\%$ confidence interval for n, we use the Imbens and Manski (2004) formula as follows.

$$[\widehat{n}_l - \bar{C}_N\widehat{\tau}_l,\ \widehat{n}_u + \bar{C}_N\widehat{\tau}_u], \tag{4.3}$$

where $\Phi\left(\bar{C}_N + \frac{\widehat{n}_u - \widehat{n}_l}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)}\right) - \Phi(-\bar{C}_N) = 1-\alpha$. The length of this interval increases with the estimated efficiency bounds ς_l and ς_u, and the estimated bounds $\widehat{\psi}_l^{-1}$ and $\widehat{\psi}_u^{-1}$. Moreover, this length increases with the sample size (equivalently the population size) at a rate $\sqrt{N}$.

Imbens and Manski (2004) have shown in the coverage of their general proposed confidence interval is not too small compared to the nominal coverage. In the following, theorem, we evaluate the actual error. We present the result for the general case in theorem 4.1. In this section, the bounds and the estimated bounds are linear functions of n and N respectively, and hence, the error for this set-up is slightly different from the one in theorem 4.1.

Theorem 4.5. *Let φ_l and φ_u be any generic influence functions for the estimation of ψ_l^{-1} and ψ_u^{-1} as in theorem 4.4. Let $\Delta = \psi_u^{-1} - \psi_l^{-1}$ and $\widehat{\Delta} = \widehat{\psi}_u^{-1} - \widehat{\psi}_l^{-1}$. Let $\widehat{\tau}_u$, $\widehat{\tau}_l$, $\widetilde{\tau}_u$ and $\widetilde{\tau}_l$ be defined as above. Further, for a given n, let $0 < \underline{\tau} \leq \widetilde{\tau}_l, \widetilde{\tau}_u, \widehat{\tau}_l, \widehat{\tau}_u \leq \bar{\tau} < \infty$ where $\underline{\tau}$ and $\bar{\tau}$ are constants. Suppose the following assumption holds.*

Assumption 6. *For a given $\epsilon > 0$ and a constant c, there exists N_0 and $\upsilon > 0$ such that for all $N > N_0$*

$$\mathbb{P}\left(|\widehat{\Delta} - \Delta| > cn^{-\upsilon}\right) < \epsilon.$$

Then the finite sample coverage error of the confidence interval in equation 4.3 of containing n_u is bounded as follows

$$
\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \le n_u \le \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u\right) \\
\le\ & \frac{\sqrt{n\psi}}{\sqrt{2\pi}}\mathbb{E}\left(1\frac{|\widehat{R}_{2,l}|}{\widetilde{\tau}_l} + \frac{|\widehat{R}_{2,u}|}{\widetilde{\tau}_u}\right) + \frac{C}{\sqrt{n}\psi^{1.5}}\mathbb{E}\left(\frac{\rho_l}{\widetilde{\tau}_l^3} + \frac{\rho_u}{\widetilde{\tau}_u^3}\right) + \mathbb{1}(\Delta \ne 0)\mathbb{E}\left\{\frac{2}{\sqrt{n}}(1+2\theta)\right\} \\
&+\ \frac{c_N}{\sqrt{2\pi}}\left\{\mathbb{E}\left(1 - \frac{\sqrt{N}\widehat{\tau}_u}{\sqrt{n\psi}\widetilde{\tau}_u}\right) + \mathbb{E}\left(1 - \frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\right\} \\
&+\ \sqrt{\frac{2}{\pi}}\mathbb{E}\left\{\frac{(N\widehat{\tau}_l^2) \vee (n\psi\widetilde{\tau}_l^2)}{n\psi(\Delta \wedge |\Delta - \widehat{R}_{2,l}|)^2}\left(1 - \frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{\Delta}{\sqrt{N}\widehat{\tau}_l}\right\} \\
&+\ \mathbb{1}(\Delta \ne 0)\ \mathbb{E}\left\{\left(\frac{1}{\sqrt{N}\widehat{\tau}_l} - \frac{1}{\sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{\sqrt{2}(\sqrt{N}\widehat{\tau}_l \vee \sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l)^2}{\sqrt{\pi}n\psi\Delta}\right\} \\
&+\ \mathbb{1}(\Delta \ne 0)\frac{2\sqrt{3}}{\alpha\sqrt{\pi}\psi^2 n^{1.5}}\mathbb{E}\left[\frac{|N - n\psi|\widehat{\Delta} + \psi c n^{1-\upsilon}}{}\ \frac{(\widetilde{\tau}_u \vee \widetilde{\tau}_l)^2}{\Delta^3}\right] \\
&+\ \mathbb{1}(\Delta \ne 0)\frac{2\sqrt{3}}{\alpha\sqrt{\pi\psi}}\mathbb{E}\left[\frac{\{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l) - \sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)\}}{(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\ \frac{(\widehat{\tau}_u \vee \widehat{\tau}_l)^2}{N\widehat{\Delta}^2} \times \mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} > 1\right)\right],
\end{aligned}
$$

where $\rho_l = \mathbb{E}\left[|\mathbb{1}(\mathbf{Y} \ne \mathbf{0})(\widehat{\phi}_l - \mathbb{Q}\widehat{\phi}_l) + \{\mathbb{1}(\mathbf{Y} \ne \mathbf{0}) - \psi\}(\widehat{R}_{2,l} + \psi_l^{-1})|^3 \middle| \mathbf{Z}^n\right]$, $\rho_u = \mathbb{E}\left[|\mathbb{1}(\mathbf{Y} \ne \mathbf{0})(\widehat{\phi}_u - \mathbb{Q}\widehat{\phi}_u) + \{\mathbb{1}(\mathbf{Y} \ne \mathbf{0}) - \psi\}(\widehat{R}_{2,u} + \psi_u^{-1})|^3 \middle| \mathbf{Z}^n\right]$, *and* θ *is the maximum value of the density of* $(\widehat{n}_u - n_u)/(\sqrt{N}\widehat{\tau}_u)$ *and* $(\widehat{n}_l - n_l)/(\sqrt{N}\widehat{\tau}_l)$, *and* C *is the Berry-Esseen constant.*

The above theorem says that the proposed confidence interval in 4.3 contains the true upper bound n_u with probability at least as large as $1-\alpha$ minus some error terms. A similar result follows for the lower bound n_l. Hence, the proposed confidence interval contains the true population size n with the same guarantee. The first term in the error bound above summarizes the second order error term. The second term comes from the Berry-Esseen normal approximation. The fourth and the fifth terms summarizes the bias in the estimated variance and/or standard deviation. These terms are large if the estimated standard deviations are smaller than the true standard deviations. The third and the remaining terms are all of order $\frac{1}{\sqrt{n}}$.

This error is sufficiently small when $\mathbb{E}|\widehat{R}_{2,l}|$ and $\mathbb{E}|\widehat{R}_{2,u}|$ are small. The details can be found in the discussion following theorem 4.3 in section 4.4.2.

Remark 26. *The coverage error presented in the above theorem is valid as long as* Δ *and* $\widehat{\Delta}$ *are bounded away from zero and depend on neither* n *nor* N*.*

The simplified large sample coverage error is presented in the following corollary.

Corollary 4.5.1. *The error in coverage is bounded as follows for a sufficiently large n.*

$$1 - \alpha - \mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \leq n \leq \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u\right) \lesssim \frac{1}{\sqrt{n}} + \sqrt{n}\mathbb{E}(|\widehat{R}_{2,l}| + |\widehat{R}_{2,u}|).$$

The above error decreases and n increases if the second order remainder terms are sufficiently small.

4.6 Simulation & Application

In this section, we check the performance of the proposed method in a simulated set-up and then apply it to real data.

4.6.1 Estimation in a simulated set-up

We have used a simulation set-up similar to the one in Das et al. (2021), but we added some dependence between the two lists. The true risk ratio function is approximately 1.2. Below is the simulation set-up.

$$\begin{aligned}
&X \sim Normal(2,1)\\
&\mathbb{P}(Y_1 = 1 \mid Y_2 = 1,\, X = x) = \text{expit}(-3.214 + 0.5x)\\
&\mathbb{P}(Y_1 = 1 \mid Y_2 = 0,\, X = x) = \frac{\text{expit}(-3.214 + 0.5x)}{1.2}\\
&\mathbb{P}(Y_2 = 1 \mid X = x) = \text{expit}(-3.214 + 0.3x).
\end{aligned}$$

In this set-up, the true capture probability ψ is approximately 0.5. One can change the intercept term to obtain a different capture probability. Also, instead of estimating the nuisance functions, we simulate the estimates by perturbing the true nuisance functions with controlled as follows.

$$\widehat{q}_j(x) = \text{expit}[\text{logit}\{q_j(x)\} + \epsilon_j]$$

where the errors ϵ_j are simulated from $\mathcal{N}(n^{-\alpha}, n^{-2\alpha})$, where $\alpha \in \{0.1, 0.25, 0.5\}$ is the error rate. This set-up allows us to study the robustness of our estimator against the error in the estimation of the nuisance functions. Figure 4.1 shows the empirical coverage of the estimated 95% confidence intervals for n from 500 iterations. The true population size is 5000.

We see in figure 4.1, that the estimate confidence intervals have coverage close to the target level for a wider range of ω when one uses the proposed method for slower error rates. The coverage of the plug-in and the proposed methods are comparable for the parametric error rate which is $n^{-0.5}$.

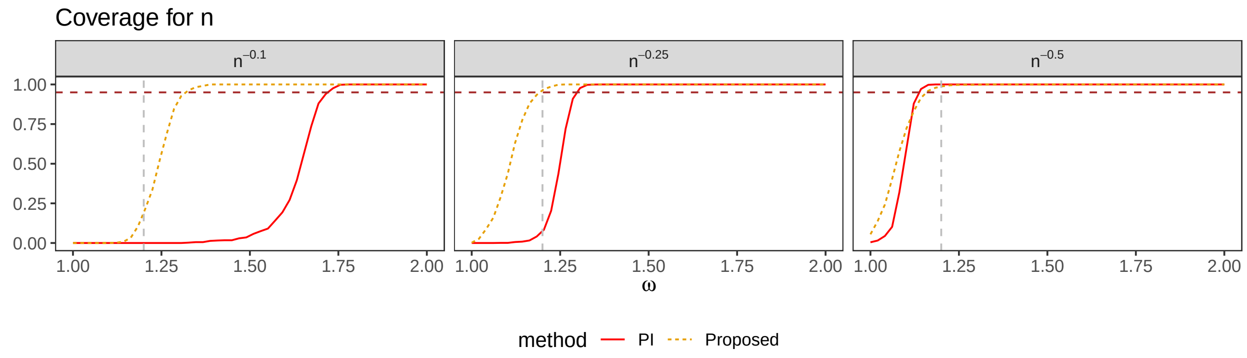

Figure 4.1: Performance plot of capture probability and total population size as a function of the risk ratio bound ω. The plots show the empirical coverage of n for different error rates in the estimation of the q-probabilities. The true risk ratio is approximately 1.2 across all the covariates. The red and the yellow lines mark the plug-in and the proposed methods. The horizontal line is the target coverage, i.e., 0.95 and the vertical line marks the true risk ratio.

4.6.2 Application on real data

For the real data application, we apply our method on the Peru Internal Armed Conflict Data from 1980-2000 (Ball et al., 2003). This dataset mainly consists of 24,692 documented victims of the war, along with some demographic and geographic covariate information. The data comes from three sources, i.e., three lists. We combined two of these lists after comparing their demographic distribution. More details are available in Das et al. (2021). In this section, we are interested in estimating the confidence interval for various levels of deviation ω from the conditional independence assumption. Figure 4.2 shows the estimates lower and upper bounds and the 95% confidence intervals for deviation upto 1.25. The exact true risk ratio is usually not known. One can use calibration on the observed data to get an approximate range for the risk ratio and select ω accordingly.

4.7 Discussion

In this paper, we have considered a sensitivity analysis approach to estimate population size in the capture-recapture set-up. Das et al. (2021) presented inference when two lists are independent conditional on covariate information. We consider a relaxation of this assumption and the degree of relaxation is the sensitivity parameter. We quantify the dependence using the risk ratio between the two lists. One can also obtain equivalent results using the odds ratio in a similar manner. We present flexible bound parameter estimates for the target parameters under a margin condition. We also present the efficient influence functions, the efficiency bounds and discuss that the proposed estimators achieve these bounds. We further construct confidence intervals using Imbens and Manski (2004) formula and present the finite sample coverage

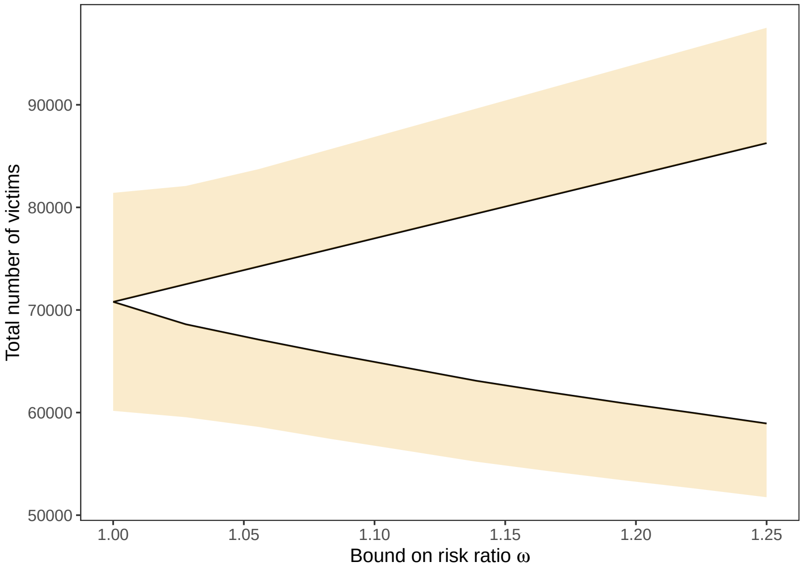

Figure 4.2: The 95% confidence interval for the total number of victims in the Peru Internal Armed Conflict of 1980-2000 as a function of the risk ratio bound ω.

guarantees for a general case. Finally, we apply the proposed approach to construct confidence intervals for the number of victims of the Peru Internal Armed Conflict 1980-2000.

Chapter 5

Conclusions

Population size estimation has applications in many areas. Hence, it is crucial to use reliable methods to avoid oversight. Existing methods often have to choose between flexible estimation and achieving the fast $\sqrt{n}$-rates. Moreover, selecting a set-up that is not too restrictive for real data is crucial.

In this thesis, we have presented a nonparametric approach for population size estimation in a capture-recapture set-up. We have discussed three problem set-ups, (i) estimation under conditional independence, (ii) user-friendly software for real data implementation, and (iii) estimation under a relaxed assumption that guarantees only partial identification for the target parameters.

In chapter 2, we have presented an estimator under a conditional independence identifiability assumption, and we have shown that the proposed approach is efficient even under a flexible set-up. Further, the proposed approach is doubly robust against errors in the estimation and is approximately normal. We also presented a general formula to construct confidence intervals and evaluated the finite sample coverage error. Finally, we estimated the size of the victim population and sub-populations on the Peru Internal Armed Conflict of 1980-2000.

In chapter 3, we presented the R package drpop that calculates the doubly robust population size estimated from the input data. We illustrated the use of the package under various set-ups with examples. Also, we presented some comparison against existing packages to demonstrate that the drpop package works.

In chapter 4, we considered a milder (in comparison to the assumption used in the previous chapters) assumption that the dependence between two lists conditioned on the covariates is bounded. We appointed a sensitivity approach for this and presented flexible estimators. We further presented confidence interval formulas for the population size and evaluated the finite sample coverage guarantees for a general confidence interval. We applied this method on the Peru data to evaluate the 95% interval as a function of the sensitivity parameter.

This thesis focuses on flexible and efficient estimation under a given lack of dependence assumption using a pair of lists. When there are more than two lists, it is often of interest to utilize the extra structure of the lists, if any. One immediate extension is when the conditional independent list pair is not known. One can use a partial identification approach in this case to construct a confidence interval, as in chapter 4. Another extension is when more than one list pair satisfies the conditional independence assumption. It can be an important future work to study the advantages of this set-up and potentially improve the proposed estimator. Considering the lack of dependence assumption in chapter 4 on the risk ratio, a variation of this assumption can be to bound the risk ratio more flexibly than with strict point-wise bounds.

Bibliography

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635. 4, 7, 53

Alho, J. M., Mulry, M. H., Wurdeman, K., and Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88(423):1130–1136. 8, 28, 32

Bailey, N. T. (1952). Improvements in the interpretation of recapture data. *The Journal of Animal Ecology*, pages 120–127. 31

Baillargeon, S. and Rivest, L.-P. (2007). Rcapture: Loglinear models for capture-recapture in r. *Journal of Statistical Software, Articles*, 19(5):1–31. 28, 48, 49

Ball, P., Asher, J., Sulmont, D., and Manrique, D. (2003). How many peruvians have died. *Washington, DC: American Association for the Advancement of Science*. xvii, 1, 3, 21, 22, 69, 98, 100, 101

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore. 2, 4, 9, 10, 12, 13, 34, 60, 62, 86

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393. 13, 34

Bolker, B. M. (2008). *Ecological Models and Data in R*. Princeton University Press. 52

Bonvini, M. and Kennedy, E. H. (2020). Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–31. 116

Brenner, H. (1995). Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology*, pages 42–48. 8

Breslow, N. E., Robins, J. M., and Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6(3):447–455. 5

Bunge, J. A. (2013). A survey of software for fitting capture–recapture models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):114–120. 29

Burnham, K. P. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60(5):927–936. 3, 7, 27, 32

Carothers, A. (1973). The effects of unequal catchability on jolly-seber estimates. *Biometrics*, pages 79–100. 3, 27

Chan, L., Silverman, B. W., and Vincent, K. (2020). Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists. *Journal of the American Statistical Association*, pages 1–10. 53

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791. 7, 32, 53

Chao, A. (2014). Capture-recapture for human populations. *Wiley StatsRef: Statistics Reference Online*, pages 1–16. 28

Chao, A., Tsay, P., Lin, S.-H., Shau, W.-Y., and Chao, D.-Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in medicine*, 20(20):3123–3157. 28

Chen, S. X. and Lloyd, C. J. (2000). A nonparametric approach to the analysis of two-stage mark-recapture experiments. *Biometrika*, 87(3):633–649. 4

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65. 13

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. 12

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317. 37

Choquet, R., Reboulet, A.-M., Pradel, R., Gimenez, O., and Lebreton, J.-D. (2004). M–surge: New software specifically designed for multistate capture–recapture models. *Animal biodiversity and conservation*, pages 207–215. 28

Cormack, R. (1985). Examples of the use of glim to analyse capture-recapture studies. In *Statistics in ornithology*, pages 243–273. Springer. 28

Cormack, R. and Jupp, P. (1991). Inference for poisson and multinomial models for capture-recapture experiments. *Biometrika*, 78(4):911–916. 28

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, pages 395–413. 28

Crawley, M. J. (1993). Glim for ecologists. 52

Darroch, J. N. (1958). The multiple-recapture census, i. estimation of a closed population. *Biometrika*, 45:343–349. 3, 7, 12, 27, 31

Das, M., Kennedy, E. H., and Jewell, N. P. (2021). Doubly robust capture-recapture methods for estimating population size. *arXiv preprint arXiv:2104.14091*. 27, 28, 29, 31, 33, 34, 35, 36, 51, 52, 53, 54, 55, 58, 64, 68, 69, 109, 119

Evans, M. A., Kim, H.-M., and O'Brien, T. E. (1996). An application of profile-likelihood based confidence interval to capture: recapture estimators. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 131–140. 19

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23. 14

Fienberg, S. E. (1972). The multiple recapture census for closed popu- lations and incomplete 2^k contingency tables. *Biometrika*, 59:591–603. 3, 7, 9, 12, 27, 32

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). Classical multilevel and bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3):383–405. 1, 3

Friedrich, K. O. (1989). A berry-esseen bound for functions of independent random variables. *The Annals of Statistics*, pages 170–183. 16

Frischer, M., Leyland, A., Cormack, R., Goldberg, D. J., Bloor, M., Green, S. T., Taylor, A., Covell, R., McKeganey, N., and Platt, S. (1993). Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in glasgow, scotland. *American Journal of Epidemiology*, 138(3):170–181. 28

Gimenez, O., Choquet, R., Lamor, L., Scofield, P., Fletcher, D., Lebreton, J.-D., and Pradel, R. (2005). Efficient profile-likelihood confidence intervals for capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2):184–196. 9

Gimenez, O., Covas, R., Brown, C. R., Anderson, M. D., Brown, M. B., and Lenormand, T. (2006). Nonparametric estimation of natural selection on a quantitative trait using mark-recapture data. *Evolution*, 60(3):460–466. 52

Goudie, I. B. and Goudie, M. (2007). Who captures the marks for the petersen estimator? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(3):825–839. 3, 31

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media. 14

Hald, A. (2003). *A History of Probability and Statistics and Their Applications Before 1750*, volume 501. John Wiley & Sons. 1, 3, 31

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393. 84, 114

Hernan, M. A. and Robins, J. M. (2019). *Causal Inference*. CRC Boca Raton, FL. 7

Hook, E. B. and Regal, R. R. (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of clinical epidemiology*, 52(10):917–26. 27

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140. 4, 5, 7, 8, 27, 28, 32, 53

Huggins, R. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, pages 725–732. 28

Huggins, R. (2001). A note on the difficulties associated with the analysis of capture–recapture experiments with heterogeneous capture probabilities. *Statistics & probability letters*, 54(2):147–152. 27

Huggins, R. and Hwang, W.-H. (2007). Non-parametric estimation of population size from capture–recapture data when the capture probability depends on a covariate. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(4):429–443. 4, 7, 12, 28, 32

Huggins, R. and Hwang, W.-H. (2011). A review of the use of conditional likelihood in capture-recapture experiments. *International Statistical Review*, 79(3):385–400. 4, 32

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857. 2, 54, 55, 56, 57, 64, 66, 69, 103, 106, 121, 128

Jolly, G. and Dickson, J. (1983). The problem of unequal catchability in mark-recapture estimation of small mammal populations. *Canadian Journal of Zoology*, 61(4):922–927. 31

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer. 2, 4, 10, 12, 13, 34, 60, 62

Kennedy, E. H. (2020). Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics*, 16(1). 86

Kennedy, E. H., Balakrishnan, S., G'Sell, M., et al. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030. 116

Krebs, C. J. et al. (2014). Ecological methodology. Technical report, Harper & Row New York. 1, 31

Kurtz, Z. T. (2018). Local log-linear models for capture-recapture. 4

Laake, J. and Rexstad, E. (2008). Rmark–an alternative approach to building linear models in mark. *Program MARK: a gentle introduction*, pages C1–C113. 28

Laake, J. L., Johnson, D. S., and Conn, P. B. (2013). marked: an r package for maximum likelihood and markov chain monte carlo analysis of capture–recapture data. *Methods in Ecology and Evolution*, 4(9):885–890. 28

Lee, S.-M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, pages 88–97. 3, 27

Lincoln, F. C. (1930). *Calculating waterfowl abundance on the basis of banding returns*. Number 118. US Department of Agriculture. 3, 7

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130. 3, 27

Lum, K. and Ball, P. (2015). Estimating undocumented homicides with two lists and list dependence. *Human Rights Data Analysis Group*. 1

Manrique-Vallier, D. (2016). Bayesian population size estimation using dirichlet process mixtures. *Biometrics*, 72(4):1246–1254. 4

Manrique-Vallier, D. and Ball, P. (2019). Reality and risk: A refutation of s. rendon's analysis of the peruvian truth and reconciliation commission's conflict mortality study. *Research & Politics*, 6(1):2053168019835628. 22

Manrique-Vallier, D., Ball, P., and Sulmont, D. (2019). Estimating the number of fatal victims of the peruvian internal armed conflict, 1980-2000: an application of modern multi-list capture-recapture techniques. *arXiv preprint arXiv:1906.04763*. 1, 22

McClintock, B. T. (2015). multimark: an r package for analysis of capture–recapture data consisting of multiple "noninvasive" marks. *Ecology and evolution*, 5(21):4920–4931. 28

McDonald, T., Regehr, E., Manly, B., Bromaghin, J., and McDonald, M. T. (2018). Package 'mra'. 28

Mises, R. v. (1947). On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348. 84, 114

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135. 53

PERÚ, G. (2014). Geo gps perú: Base de datos perú–shapefile–*. shp–minam–ign–límites políticos. 98

Petersen, C. G. J. (1896). The yearly immigration of young plaice in the limfjord from the german sea. *Rept. Danish Biol. Sta.*, 6:1–48. 1, 3, 7, 27, 31, 53

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2):434–442. 7, 32, 53

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: an overview. *Journal of Applied Statistics*, 29(1-4):85–102. 4, 5, 28

Pollock, K. H., Nichols, J. D., Brownie, C., and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife monographs*, pages 3–97. 7, 8, 32, 53

Qin, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer. 5

Rendon, S. (2019a). Capturing correctly: a reanalysis of the indirect capture-recapture methods in the peruvian truth and reconciliation commission. *Research & Politics*, 6(1):2053168018820375. 21, 22, 98

Rendon, S. (2019b). *Replication Data for: "Capturing Correctly: A Reanalysis of the Indirect Capture-recapture Methods in the Peruvian Truth and Reconciliation Commission"*. Harvard Dataverse. 98

Rivest, L.-P. and Baillargeon, S. (2007). Applications and extensions of chao's moment estimator for the size of a closed population. *Biometrics*, 63(4):999–1006. 28

Rivest, L.-P. and Daigle, G. (2004). Loglinear models for the robust design in mark–recapture experiments. *Biometrics*, 60(1):100–107. 28

Rivest, L.-P. and Lévesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *Canadian Journal of Statistics*, 29(4):555–572. 28

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics. 4, 13, 37

Schluter, D. (1988). Estimating the form of natural selection on a quantitative trait. *Evolution*, 42(5):849–861. 52

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. *The American Mathematical Monthly*, 45(6):348–352. 3, 27, 31

Seber, G. A. F. et al. (1982). *The Estimation of Animal Abundance and Related Parameters*, volume 8. Blackburn press Caldwell, New Jersey. 31

Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44(245):101–115. 7, 32, 53

Stoklosa, J. and Huggins, R. M. (2012). A robust p-spline approach to closed population capture–recapture models with time dependence and heterogeneity. *Computational Statistics & Data Analysis*, 56(2):408–417. 4, 28

Tilling, K. (2001). Capture-recapture methods- useful or misleading? 8, 27

Tilling, K. and Sterne, J. A. (1999). Capture-recapture models including covariate effects. *American journal of epidemiology*, 149(4):392–400. 4, 5, 7, 8, 12, 20, 27, 32, 34, 53

Tsiatis, A. (2006). Semiparametric theory and missing data. *New York*. 10, 12, 13, 34, 60, 62, 86

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media. 14

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media. 2, 4, 10, 12, 13, 34, 60, 86

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media. 17, 36

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1). 17, 29, 34, 36

van der Vaart, A. (2002a). Part iii: Semiparameric statistics. *Lectures on Probability Theory and Statistics*, pages 331–457. 10, 12, 13, 34, 60

van der Vaart, A. (2014). Higher order tangent spaces and influence functions. *Statistical Science*, pages 679–686. 4

van der Vaart, A. W. (2002b). Semiparametric statistics. *Lecture Notes in Math.*, (1781). 10, 11, 34, 62, 86

Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J., and Robins, J. M. (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, 103(484):1693–1704. 15

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, 13(1):178–203. 5

White, G. C. and Burnham, K. P. (1999). Program mark: Survival estimation from populations of marked animals. *Bird study*, 46(sup1):S120–S139. 28

White, G. C., Burnham, K. P., and Anderson, D. R. (2001). Advanced features of program mark. In *Wildlife, land, and people: priorities for the 21st century. Proceedings of the second international wildlife management congress. The Wildlife Society, Bethesda, Maryland, USA*, pages 368–377. 28

Wu, Y., Yang, P., et al. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883. 12

Yee, T. W. and Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of vegetation science*, 2(5):587–602. 52

Yee, T. W., Stoklosa, J., and Huggins, R. M. (2015). The vgam package for capture-recapture data using the conditional likelihood. *Journal of Statistical Software*, 65(1):1–33. 4, 28, 29, 52

Yip, P. S., Wan, E. C., and Chan, K. S. (2001). A unified approach for estimating population size in capture-recapture studies with arbitrary removals. *Journal of agricultural, biological, and environmental statistics*, 6(2):183–194. 4, 28

You, Y., van der Laan, M., Collender, P., Cheng, Q., Hubbard, A., Jewell, N. P., Hu, Z. T., Mejia, R., and Remais, J. (2021). Estimation of population size based on capture recapture designs and evaluation of the estimation reliability. *arXiv preprint arXiv:2105.05373*. 7, 27, 32, 53

Zheng, W. and van der Laan, M. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation, uc berkeley division of biostatistics working paper series. 13, 37

Zwane, E. and van der Heijden, P. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling*, 5(1):39–52. 4

Appendix

Appendix A

Appendix for chapter 1

A.1 Proofs of theorems and results

Proof of proposition 1. The goal is to express ψ in terms of the observed data distribution $\mathbb{Q}$. By definition $\psi = \mathbb{P}(\mathbf{Y} \neq \mathbf{0})$ and $\gamma(\mathbf{x}) = \mathbb{P}(\mathbf{Y} \neq \mathbf{0} | \mathbf{X} = \mathbf{x})$.

First we show that ψ is the harmonic mean of $\gamma(\mathbf{x})$. It is easy to see that

$$\frac{\gamma(\mathbf{x})}{\psi} = \frac{\mathbb{P}(\mathbf{Y} \neq \mathbf{0} | \mathbf{X} = \mathbf{x})}{\mathbb{P}(\mathbf{Y} \neq \mathbf{0})} = \frac{\mathbb{P}(\mathbf{Y} \neq \mathbf{0},\, \mathbf{X} = \mathbf{x})}{\mathbb{P}(\mathbf{X} = \mathbf{x})\mathbb{P}(\mathbf{Y} \neq \mathbf{0})} = \frac{\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{Y} \neq \mathbf{0})}{\mathbb{P}(\mathbf{X} = \mathbf{x})}.$$

Thus,

$$\begin{aligned} \int \frac{\mathbb{P}(\mathbf{X} = \mathbf{x})}{\psi} d\mathbf{x} &= \int \frac{\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{Y} \neq \mathbf{0})}{\gamma(\mathbf{x})} d\mathbf{x} \\ \text{or, } \frac{1}{\psi} &= \int \frac{\mathbb{Q}(\mathbf{X} = \mathbf{x})}{\gamma(\mathbf{x})} d\mathbf{x}, \text{ since } \mathbb{Q}(\mathbf{Z}) = \mathbb{P}(\mathbf{Z} | \mathbf{Y} \neq \mathbf{0}). \end{aligned}$$

We have shown a that the capture probability ψ is the harmonic mean of the conditional capture probabilities $\gamma(\mathbf{x})$ under the observed data distribution $\mathbb{Q}$.

Next, we express $\gamma(\mathbf{x})$ in terms of the q-probabilities. Under assumption 1, for any fixed covariate value $\mathbf{x}$,

$$\mathbb{P}(Y_1 = 1,\, Y_2 = 1 | \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_1 = 1 | \mathbf{X} = \mathbf{x})\mathbb{P}(Y_2 = 1 | \mathbf{X} = \mathbf{x}).$$

For a capture history $\mathbf{y} \neq \mathbf{0}$,

$$\begin{aligned}\mathbb{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x}) &= \frac{\mathbb{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x})\mathbb{P}(\mathbf{Y}\neq\mathbf{0}|\mathbf{X}=\mathbf{x})}{\mathbb{P}(\mathbf{Y}\neq\mathbf{0}|\mathbf{X}=\mathbf{x})} \\ &= \mathbb{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x},\mathbf{Y}\neq\mathbf{0})\gamma(\mathbf{x}) \\ &= \mathbb{Q}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x})\gamma(\mathbf{x}).\end{aligned}$$

Substituting the above relation in the assumption statement and using the notation of the q-probabilities, we get

$$\begin{aligned}&or,\ \mathbb{Q}(Y_1=1, Y_2=1|\mathbf{X}=\mathbf{x})\gamma(\mathbf{x}) = \mathbb{Q}(Y_1=1|\mathbf{X}=\mathbf{x})\mathbb{Q}(Y_2=1|\mathbf{X}=\mathbf{x})\gamma(\mathbf{x})^2 \\ &or,\ \gamma(\mathbf{x}) = \frac{q_{12}(\mathbf{x})}{q_1(\mathbf{x})q_2(\mathbf{x})}.\end{aligned}$$

□

Proof of lemma 2.0.1. There are several ways one can derive an efficient influence function and corresponding efficiency bound. Here, we first find a putative influence function in the discrete case using a standard Gateaux derivative argument. Then we show that this influence function is actually the efficient one in a general nonparametric model (continuous or discrete or mixed), by checking that the corresponding remainder term in a von Mises expansion is second-order.

To find a candidate influence function, we consider a special parametric submodel (i.e., deviation from $\mathbb{Q}$) given by $\mathbb{Q}_\epsilon$ with density $q_\epsilon = (1-\epsilon)q(\mathbf{z}) + \epsilon\bar{q}(\mathbf{z})$ where $\bar{q} = \bar{q}(\mathbf{z}) = \mathbb{1}(\mathbf{z} = \tilde{\mathbf{z}})$ is a point mass at $\mathbf{Z} = \tilde{\mathbf{z}}$, and for which the pathwise derivative

$$\frac{\partial}{\partial\epsilon}\left\{\frac{1}{\psi(\mathbb{Q}_\epsilon)}\right\}\Bigg|_{\epsilon=0}$$

actually equals the influence function (in the discrete case) Mises (1947); Hampel (1974). We also let $q_{s,\epsilon}(\mathbf{x})$ denote the analog of $q_s(\mathbf{x})$ under the submodel for $s \in \{1, 2, 12\}$, e.g., the marginal density for $\mathbf{X}$ under $\mathbb{Q}_\epsilon$ is

$$q_\epsilon(\mathbf{x}) = \sum_{\mathbf{y}} q_\epsilon(\mathbf{z}) = (1-\epsilon)q(\mathbf{x}) + \epsilon\mathbb{1}(\mathbf{x} = \tilde{\mathbf{x}}).$$

Now the above pathwise derivative equals

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\left\{\frac{1}{\psi(\mathbb{Q}_\epsilon)}\right\}\Bigg|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\int \frac{q_{2,\epsilon}(\mathbf{x})q_{1,\epsilon}(\mathbf{x})}{q_{12,\epsilon}(\mathbf{x})}q_\epsilon(\mathbf{x})d\mathbf{x}\Bigg|_{\epsilon=0} \\
&= \int \frac{\partial}{\partial\epsilon}\left\{\frac{q_{2,\epsilon}(\mathbf{x})q_{1,\epsilon}(\mathbf{x})}{q_{12,\epsilon}(\mathbf{x})}q_\epsilon(\mathbf{x})\right\}d\mathbf{x}\Bigg|_{\epsilon=0} \\
&= \int \frac{q_{2,\epsilon}(\mathbf{x})q_{1,\epsilon}(\mathbf{x})}{q_{12,\epsilon}(\mathbf{x})}q_\epsilon(\mathbf{x})\left\{\frac{q'_{1,\epsilon}(\mathbf{x})}{q_{1,\epsilon}(\mathbf{x})}+\frac{q'_{2,\epsilon}(\mathbf{x})}{q_{2,\epsilon}(\mathbf{x})}-\frac{q'_{12,\epsilon}(\mathbf{x})}{q_{12,\epsilon}(\mathbf{x})}+\frac{q'_\epsilon(\mathbf{x})}{q_\epsilon(\mathbf{x})}\right\}d\mathbf{x}\Bigg|_{\epsilon=0}.
\end{aligned}$$

where the last line follows by the product rule. For the discrete case, we use the notation of the integral to denote summation over $\mathbf{x}$.

For the derivatives appearing above, by the definition of $\bar{q}$, we have $q'_\epsilon(\mathbf{x}) = \frac{\partial}{\partial\epsilon}q_\epsilon(\mathbf{x}) = \mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}}) - q(\mathbf{x})$. Similarly, by using derivative of product rule for $q'_{1,\epsilon}(\mathbf{x})$, we get

$$\begin{aligned}
q'_{1,\epsilon}(\mathbf{x}) = \frac{\partial}{\partial\epsilon}q_{1,\epsilon}(\mathbf{x}) &= \frac{\partial}{\partial\epsilon}\frac{\mathbb{Q}_\epsilon(Y_1=1,\mathbf{X}=\mathbf{x})}{q_\epsilon(\mathbf{x})} \\
&= \frac{\partial}{\partial\epsilon}\frac{(1-\epsilon)\mathbb{Q}(Y_1=1,\mathbf{X}=\mathbf{x})+\epsilon\mathbb{1}(\tilde{Y}_1=1,\mathbf{x}=\tilde{\mathbf{x}})}{(1-\epsilon)q(\mathbf{x})+\epsilon\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})}, \text{ where } \tilde{Y}_1=\mathbb{1}(\tilde{y}_1=1) \\
&= \frac{-\mathbb{Q}(Y_1=1,\mathbf{X}=\mathbf{x})+\mathbb{1}(\tilde{Y}_1=1,\mathbf{x}=\tilde{\mathbf{x}})}{(1-\epsilon)q(\mathbf{x})+\epsilon\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})} \\
&\quad - \frac{(1-\epsilon)\mathbb{Q}(Y_1=1,\mathbf{X}=\mathbf{x})+\epsilon\mathbb{1}(\tilde{Y}_1=1,\mathbf{x}=\tilde{\mathbf{x}})}{\{(1-\epsilon)q(\mathbf{x})+\epsilon\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})\}^2}q'_\epsilon(\mathbf{x}).
\end{aligned}$$

The last step follows from the product rule of derivatives. Finally, setting $\epsilon = 0$, we get $q'_{1\epsilon}(\mathbf{x})|_{\epsilon=0} = \frac{\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})}{q(\mathbf{x})}\{\tilde{Y}_1 - q_1(\mathbf{x})\}$. The derivatives for $q_{2,\epsilon}$ and $q_{12,\epsilon}$ follow similarly.

Thus, combining the above results and using the discrete nature of the distribution, we get

$$\begin{aligned}
\phi &= \frac{\partial}{\partial\epsilon}\left\{\frac{1}{\psi(\mathbb{Q}_\epsilon)}\right\}\Bigg|_{\epsilon=0} \\
&= \sum_{\mathbf{x}}\frac{q_1(\mathbf{x})q_2(\mathbf{x})}{q_{12}(\mathbf{x})}\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})\left\{\frac{\tilde{Y}_1-q_1(\mathbf{x})}{q_1(\mathbf{x})}+\frac{\tilde{Y}_2-q_2(\mathbf{x})}{q_2(\mathbf{x})}-\frac{\tilde{Y}_1\tilde{Y}_2-q_{12}(\mathbf{x})}{q_{12}(\mathbf{x})}\right\} \\
&\quad + \sum_{\mathbf{x}}\frac{q_1(\mathbf{x})q_2(\mathbf{x})}{q_{12}(\mathbf{x})}\{\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})-q(\mathbf{x})\} \\
&= \frac{1}{\gamma(\tilde{\mathbf{x}})}\left\{\frac{\tilde{Y}_1}{q_2(\tilde{\mathbf{x}})}+\frac{\tilde{Y}_2}{q_2(\tilde{\mathbf{x}})}-\frac{\tilde{Y}_1\tilde{Y}_2}{q_{12}(\tilde{\mathbf{x}})}\right\}-\frac{1}{\psi}, \text{ using the definition of } \gamma(\mathbf{x}).
\end{aligned}$$

Now that we have a candidate influence function that is valid in a discrete model, we evaluate the remainder term R_2 in a general submodel, to show that it is actually the efficient influence function in the general case as well.

Letting $\bar{\psi} = \psi(\bar{\mathbb{Q}})$ for a generic distribution $\bar{\mathbb{Q}}$, the remainder of the so-called von Mises expansion (Bickel et al., 1993; van der Laan and Robins, 2003) is then given by:

$$\begin{aligned}
R_2(\mathbb{Q}, \bar{\mathbb{Q}}) &\equiv \frac{1}{\bar{\psi}} - \frac{1}{\psi} + \int \phi(\mathbf{z}; \bar{\mathbb{Q}}) d\mathbb{Q} \\
&= \int \left[\frac{1}{\bar{\gamma}(\mathbf{x})} \left\{ \frac{Y_1}{\bar{q}_1(\mathbf{x})} + \frac{Y_2}{\bar{q}_2(\mathbf{x})} - \frac{Y_1 Y_2}{\bar{q}_{12}(\mathbf{x})} \right\} - \frac{1}{\bar{\psi}} + \frac{1}{\bar{\psi}} - \frac{1}{\psi} \right] d\mathbb{Q}(\mathbf{x}) \\
&= \int \left[\frac{1}{\bar{\gamma}(\mathbf{x})} \left\{ \frac{q_1(\mathbf{x})}{\bar{q}_1(\mathbf{x})} + \frac{q_2(\mathbf{x})}{\bar{q}_2(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} \right\} - \frac{1}{\gamma(\mathbf{x})} \right] d\mathbb{Q}(\mathbf{x}) \\
&= \int \left\{ \frac{q_1(\mathbf{x})\bar{q}_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} + \frac{\bar{q}_1(\mathbf{x})q_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{\gamma}(\mathbf{x})\bar{q}_{12}(\mathbf{x})} - \frac{1}{\gamma(\mathbf{x})} \right\} d\mathbb{Q}(\mathbf{x}) \\
&= \int \left\{ \frac{q_1(\mathbf{x})\bar{q}_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} + \frac{\bar{q}_1(\mathbf{x})q_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} - \frac{\bar{q}_1(\mathbf{x})\bar{q}_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} - \frac{q_1(\mathbf{x})q_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} \right. \\
&\qquad \left. + \frac{\bar{q}_1(\mathbf{x})\bar{q}_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} + \frac{q_1(\mathbf{x})q_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{\gamma}(\mathbf{x})\bar{q}_{12}(\mathbf{x})} - \frac{1}{\gamma(\mathbf{x})} \right\} d\mathbb{Q}(\mathbf{x}) \\
&= \int \frac{1}{\bar{q}_{12}(\mathbf{x})} \Bigg[\{q_1(\mathbf{x}) - \bar{q}_1(\mathbf{x})\} \{\bar{q}_2(\mathbf{x}) - q_2(\mathbf{x})\} \\
&\qquad + \{q_{12}(\mathbf{x}) - \bar{q}_{12}(\mathbf{x})\} \left\{ \frac{1}{\gamma(\mathbf{x})} - \frac{1}{\bar{\gamma}(\mathbf{x})} \right\} \Bigg] d\mathbb{Q}(\mathbf{x}).
\end{aligned}$$

The above shows that the remainder of the von Mises expansion is second-order, i.e., involving products of differences between components of $\mathbb{Q}$ and $\bar{\mathbb{Q}}$. This fact implies the more general pathwise differentiability condition that

$$\frac{\partial}{\partial \epsilon} \left\{ \frac{1}{\psi(\mathbb{Q}_\epsilon)} \right\} \Bigg|_{\epsilon=0} = \int \phi \frac{\partial}{\partial \epsilon} \left\{ log\ q_\epsilon(\mathbf{z}) \Big|_{\epsilon=0} \right\} d\mathbb{Q}.$$

for any smooth parametric submodel $\mathbb{Q}_\epsilon$ (Kennedy, 2020). Thus, the candidate influence function satisfies the above pathwise differentiability condition and hence, is an efficient influence function. Moreover, since the model is non-parametric, ϕ is the only efficient influence function (Bickel et al., 1993; Tsiatis, 2006; van der Vaart, 2002b). $\square$

***Proof of Theorem* 2.1.** We will evaluate the variance of the efficient influence function $\phi \equiv \phi(\mathbf{Z};\mathbb{Q})$. Recall, $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. Hence, we use the law of total variance by conditioning on $\mathbf{X}$.

$$
\begin{aligned}
var(\phi) &= var\left[\frac{1}{\gamma(\mathbf{X})}\left\{\frac{Y_1}{q_1(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right\} - \frac{1}{\psi}\right] \\
&= var\left(\mathbb{E}\left[\frac{1}{\gamma(\mathbf{X})}\left\{\frac{Y_1}{q_1(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right\}\middle|\mathbf{X}\right]\right) \\
&+ \mathbb{E}\left(var\left[\frac{1}{\gamma(\mathbf{X})}\left\{\frac{Y_1}{q_1(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right\}\middle|\mathbf{X}\right]\right), \text{ by the law of total variance} \\
&= var\left\{\frac{1}{\gamma(\mathbf{X})}\right\} + \mathbb{E}\left[\frac{1}{\gamma(\mathbf{X})^2}\left\{\frac{1-q_1(\mathbf{X})}{q_1(\mathbf{X})} + \frac{1-q_2(\mathbf{X})}{q_2(\mathbf{X})} + \frac{1-q_{12}(\mathbf{X})}{q_{12}(\mathbf{X})}\right.\right. \\
&+ 2\frac{q_{12}(\mathbf{X}) - q_1(\mathbf{X})q_2(\mathbf{X})}{q_1(\mathbf{X})q_2(\mathbf{X})} - 2\,\frac{q_{12}(\mathbf{X}) - q_1(\mathbf{X})q_{12}(\mathbf{X})}{q_1(\mathbf{X})q_{12}(\mathbf{X})} - 2\,\left.\left.\frac{q_{12}(\mathbf{X}) - q_2(\mathbf{X})q_{12}(\mathbf{X})}{q_2(\mathbf{X})q_{12}(\mathbf{X})}\right\}\right] \\
&= var\left\{\frac{1}{\gamma(\mathbf{X})}\right\} + \mathbb{E}\left[\frac{1}{\gamma(\mathbf{X})^2}\left\{2\gamma(\mathbf{X}) - \frac{1}{q_1(\mathbf{X})} - \frac{1}{q_2(\mathbf{X})} + \frac{1}{q_{12}(\mathbf{X})} - 1\right\}\right].
\end{aligned}
$$

The last two equalities follow by evaluating the conditional expectation and the variance the two terms respectively and the relation that $\gamma(\mathbf{x}) = \frac{q_{12}(\mathbf{x})}{q_1(\mathbf{x})q_2(\mathbf{x})}$. By simple algebra, we can further simplify the variance as follows

$$
\begin{aligned}
var(\phi) &= var\left\{\frac{1}{\gamma(\mathbf{X})}\right\} + \mathbb{E}\left[\frac{1}{\gamma(\mathbf{X})^2}\left\{2\gamma(\mathbf{X}) - \frac{1}{q_1(\mathbf{X})} - \frac{1}{q_2(\mathbf{X})} + \frac{1}{q_{12}(\mathbf{X})} - 1\right\}\right] \\
&= var\left\{\frac{1}{\gamma(\mathbf{X})}\right\} + \mathbb{E}\left[\frac{1}{\gamma(\mathbf{X})^2}\left\{2\gamma(\mathbf{X}) - \frac{1 + q_{12}(\mathbf{X}) - q_0(\mathbf{X})}{q_1(\mathbf{X})q_2(\mathbf{X})} + \frac{1}{q_{12}(\mathbf{X})} - 1\right\}\right] \\
&= var\left\{\frac{1}{\gamma(\mathbf{X})}\right\} + \mathbb{E}\left(\frac{1}{\gamma(\mathbf{X})}\left[\left\{\frac{1}{\gamma(\mathbf{X})} - 1\right\}\left\{\frac{1}{q_{12}(\mathbf{X})} - 1\right\} + \frac{q_0(\mathbf{X})}{q_{12}(\mathbf{X})}\right]\right).
\end{aligned}
$$

Recall: $q_0(\mathbf{X}) = \mathbb{Q}(Y_1 = 0, Y_2 = 0 \mid \mathbf{Y} \neq \mathbf{0}, \mathbf{X} = \mathbf{X})$. □

***Proof of Theorem* 2.2.** Let $\mathcal{E}_N = \widehat{\psi}_{dr}^{-1} - \psi^{-1} - \mathbb{Q}_N\phi$. We can expand it as follows.

$$
\begin{aligned}
\mathcal{E}_N &= \frac{1}{\widehat{\psi}_{dr}} - \frac{1}{\psi} - \mathbb{Q}_N\phi \\
&= \frac{1}{\widehat{\psi}_{pi}} + \mathbb{Q}_N\widehat{\phi} - \frac{1}{\psi} - \mathbb{Q}_N\phi, \quad \text{using the definition in equation 2.9} \\
&= -\int \widehat{\phi}(\mathbf{z})d\mathbb{Q}(\mathbf{z}) + \widehat{R}_2 + \mathbb{Q}_N\widehat{\phi} - \mathbb{Q}_N\phi, \quad \text{using von Mises expansion of } \widehat{\psi}^{-1}.
\end{aligned}
$$

The last step follows using $\frac{1}{\widehat{\psi}} - \frac{1}{\psi} = -\int \widehat{\phi}(\mathbf{z})d\mathbb{Q}(\mathbf{z}) + \widehat{R}_2$ from the proof of lemma 2.0.1. The formula of $\widehat{R}_2$ is presented in the proof of lemma 2.0.1.

For simplicity of notation in the proof we will use $\mathbb{Q}(\cdot) = \mathbb{E}(\cdot|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n) = \int(\cdot)d\mathbb{Q}(\mathbf{z})$ to denote the observed population average function conditioned on the training sample $\mathbf{Z}^n$. Using that fact that $\mathbb{Q}\phi = 0$, we get

$$\mathcal{E}_N = \widehat{R}_2 + (\mathbb{Q}_N - \mathbb{Q})(\widehat{\phi} - \phi).$$

We want to bound $\mathcal{E}_N$ by presenting a bound on $\mathbb{E}(\mathcal{E}_N^2)$. We will show that the expected value of $\mathcal{E}$ is expected value of the error term $\widehat{R}_2$.

$$\begin{aligned}
\mathbb{E}(\mathcal{E}_N) &= \mathbb{E}\left\{\mathbb{E}\left(\mathcal{E}_N|\mathbf{Z}^n\right)\right\} \\
&= \mathbb{E}\left[\mathbb{E}\left\{\widehat{R}_2 + (\mathbb{Q}_N - \mathbb{Q})(\widehat{\phi} - \phi)\big|\mathbf{Z}^n\right\}\right] \\
&= \mathbb{E}(\widehat{R}_2) + \mathbb{E}\left[\mathbb{E}\left\{(\mathbb{Q}_N - \mathbb{Q})(\widehat{\phi} - \phi)\big|\mathbf{Z}^n\right\}\right].
\end{aligned}$$

Next, we will show that the second quantity is 0.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left\{(\mathbb{Q}_N - \mathbb{Q})(\widehat{\phi} - \phi)\big|\mathbf{Z}^n\right\}\right] &= \mathbb{E}\left[\mathbb{E}\left\{\mathbb{Q}_N(\widehat{\phi} - \phi) - \mathbb{Q}(\widehat{\phi} - \phi)\big|\mathbf{Z}^n\right\}\right] \\
&= \mathbb{E}\left\{\mathbb{E}\left(\widehat{\phi} - \phi|\mathbf{Z}^n\right) - \mathbb{E}\left(\widehat{\phi} - \phi|\mathbf{Z}^n\right)\right\} \\
&= 0.
\end{aligned}$$

The second equality follows because $\mathbb{Q}_N$ is evaluated on i.i.d. terms and $\mathbb{Q}(\widehat{\phi} - \phi) = \mathbb{E}(\widehat{\phi} - \phi|\mathbf{Z}^n)$ by definition. Next, we evaluate the variance of $\mathcal{E}_N$.

$$\begin{aligned}
var(\mathcal{E}_N) &= \mathbb{E}\{var(\mathcal{E}_N|\mathbf{Z}^n)\} + var\{\mathbb{E}(\mathcal{E}_N|\mathbf{Z}^n)\}, \text{ by the law of total variance} \\
&= \mathbb{E}\{var(\mathcal{E}_N|\mathbf{Z}^n)\} + var(\widehat{R}_2) \\
&= \mathbb{E}\left[var\left\{\widehat{R}_2 + (\mathbb{Q}_N - \mathbb{Q})(\widehat{\phi} - \phi)\Big|\mathbf{Z}^n\right\}\right] + var(\widehat{R}_2) \\
&= \mathbb{E}\left[var\left\{\mathbb{Q}_N(\widehat{\phi} - \phi)\Big|\mathbf{Z}^n\right\}\right] + var(\widehat{R}_2).
\end{aligned}$$

The last equality follows because $\widehat{R}_2$ and $\mathbb{Q}(\cdot)$ are constants when conditioned on $\mathbf{Z}^n$. Now,

$$\begin{aligned}
var\left\{\mathbb{Q}_N(\widehat{\phi} - \phi)\big|\mathbf{Z}^n\right\} &= \frac{1}{N}var(\widehat{\phi} - \phi|\mathbf{Z}^n), \text{ property of variance of sample average} \\
&\leq \frac{1}{N}\|\widehat{\phi} - \phi\|^2.
\end{aligned}$$

Thus, combining the above results we get $\mathbb{E}(\mathcal{E}_N^2) \leq \mathbb{E}(\widehat{R}_2^2) + \frac{\mathbb{E}\|\widehat{\phi} - \phi\|^2}{N}$.

Hence, by using Markov's inequality combined with the above inequality gives the required result.

$$\mathbb{P}\left(\left|\widehat{\psi}_{dr}^{-1} - \psi^{-1} - \mathbb{Q}_N\phi\right| \leq \delta\right) \geq 1 - \frac{1}{\delta^2}\mathbb{E}(\mathcal{E}_N^2).$$

$\square$

Proof of Theorem 2.3. The error in the estimation of ψ^{-1} is $\widehat{\psi}^{-1} - \psi^{-1}$. Following the proof of Theorem 2.2, it can be expressed as

$$\widehat{\psi}^{-1} - \psi^{-1} = (\mathbb{Q}_N - \mathbb{Q})\widehat{\phi} + \widehat{R}_2,$$

where $\mathbb{Q}\widehat{\phi}$ denotes $\int \widehat{\phi}(\mathbf{z})d\mathbb{Q}(\mathbf{z})$ as mentioned in the proof of Theorem 2.2.

Let $\widetilde{\sigma}^2 = var(\widehat{\phi} \mid \mathbf{Z}^n)$, and note $\widehat{\psi}^{-1} - \psi^{-1} - \widehat{R}_2 = (\mathbb{Q}_N - \mathbb{Q})\widehat{\phi}$ is a sample average of a fixed function given the training sample $\mathbf{Z}^n$. Therefore by Berry-Esseen we have for any t' and N that

$$\Phi(t') - \frac{C\rho}{\widetilde{\sigma}^3\sqrt{N}} \leq \mathbb{P}\left(\frac{\widehat{\psi}^{-1} - \psi^{-1} - \widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}} \leq t' \;\middle|\; \mathbf{Z}^n\right) \leq \Phi(t') + \frac{C\rho}{\widetilde{\sigma}^3\sqrt{N}}$$

Taking $t' = \frac{\widehat{\sigma}}{\widetilde{\sigma}}t - \frac{\widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}}$ for $\widehat{\sigma}^2 = \widehat{var}(\widehat{\phi})$ and $\rho = \mathbb{E}\left(|\widehat{\phi} - \mathbb{Q}\widehat{\phi}|^3 \middle| \mathbf{Z}^n\right)$, this implies

$$\Phi\left(\frac{\widehat{\sigma}}{\widetilde{\sigma}}t - \frac{\widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}}\right) - \frac{C\rho}{\widetilde{\sigma}^3\sqrt{N}} \leq \mathbb{P}\left(\frac{\widehat{\psi}^{-1} - \psi^{-1}}{\widehat{\sigma}/\sqrt{N}} \leq t \;\middle|\; \mathbf{Z}^n\right) \leq \Phi\left(\frac{\widehat{\sigma}}{\widetilde{\sigma}}t - \frac{\widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}}\right) + \frac{C\rho}{\widetilde{\sigma}^3\sqrt{N}}$$

Now note by the mean value theorem, for some t_n between t and $t + \left(\frac{\widehat{\sigma}}{\widetilde{\sigma}} - 1\right)t - \frac{\widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}}$, we have that

$$\begin{aligned}\left|\Phi\left(t + \left(\frac{\widehat{\sigma}}{\widetilde{\sigma}} - 1\right)t - \frac{\widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}}\right) - \Phi(t)\right| &= \left|\Phi'(t_n)\left\{\left(\frac{\widehat{\sigma}}{\widetilde{\sigma}} - 1\right)t - \frac{\widehat{R}_2}{\widetilde{\sigma}/\sqrt{N}}\right\}\right| \\ &\leq \frac{1}{\sqrt{2\pi}}\left(\left|\frac{\widehat{\sigma}}{\widetilde{\sigma}} - 1\right||t| + \frac{|\widehat{R}_2|}{\widetilde{\sigma}/\sqrt{N}}\right) \equiv \Delta_n\end{aligned}$$

where the second inequality used the fact that $\sup_t \Phi'(t) \leq 1/\sqrt{2\pi}$ and the triangle inequality.

Therefore

$$-\Delta_n - \frac{C\rho}{\widetilde{\sigma}^3\sqrt{N}} \leq \mathbb{P}\left(\frac{\widehat{\psi}^{-1} - \psi^{-1}}{\widehat{\sigma}/\sqrt{N}} \leq t \;\middle|\; \mathbf{Z}^n\right) - \Phi(t) \leq \Delta_n + \frac{C\rho}{\widetilde{\sigma}^3\sqrt{N}}$$

This implies by iterated expectation that

$$\left|\mathbb{P}\left(\frac{\widehat{\psi}^{-1} - \psi^{-1}}{\widehat{\sigma}/\sqrt{N}} \leq t\right) - \Phi(t)\right| \leq \sqrt{\frac{1}{2\pi}}\left\{\sqrt{N}\mathbb{E}\left(\frac{|\widehat{R}_2|}{\widetilde{\sigma}}\right) + |t|\mathbb{E}\left(\left|\frac{\widehat{\sigma}}{\widetilde{\sigma}} - 1\right|\right)\right\} + \frac{C}{\sqrt{N}}\mathbb{E}\left(\frac{\rho}{\widetilde{\sigma}^3}\right).$$

Therefore if $\widetilde{\sigma} \gtrsim 1$ with probability one, $\mathbb{E}(\rho) < c'$ for some $c' > 0$, $\mathbb{E}|\widetilde{\sigma} - \widehat{\sigma}| \lesssim n^{-1/2}$ and $\mathbb{E}|\widehat{R}_2| \lesssim n^{-2\beta}$, then

$$\left| \mathbb{P}\left(\frac{\widehat{\psi}^{-1} - \psi^{-1}}{\widehat{\sigma}/\sqrt{N}} \leq t \right) - \Phi(t) \right| \lesssim n^{-1/2} + n^{(1-4\beta)/2}$$

If $\beta > 1/4$, then there exists an $N = N_\epsilon$ guaranteeing that the LHS is no more than ϵ.

Bound on $\mathbb{E}|\widehat{\sigma} - \widetilde{\sigma}|$

We will show that this bound is $n^{-1/2}$. For $\widehat{\phi}$, we have defined the quantities $\widetilde{\sigma}^2 = var(\widehat{\phi} \mid \mathbf{Z}^n)$ and $\widehat{\sigma}^2 = \widehat{var}(\widehat{\phi})$. We can further expand $\widetilde{\sigma}^2$ as follows

$$\widetilde{\sigma}^2 = var(\widehat{\phi} \mid \mathbf{Z}^n) = \mathbb{E}\left(\widehat{\phi}^2 | \mathbf{Z}^n\right) - \left\{\mathbb{E}\left(\widehat{\phi} | \mathbf{Z}^n\right)\right\}^2.$$

The second quantity $\widehat{\sigma}^2$ is the unbiased estimator of the variance of $\widehat{\phi}$ given the test sample i.e. $\frac{N}{N-1}\left\{\mathbb{Q}_N \widehat{\phi}^2 - (\mathbb{Q}_N \widehat{\phi})^2\right\}$. Then by iterated expectation

$$\mathbb{E}(\widehat{\sigma}^2) = \mathbb{E}\{\mathbb{E}(\widehat{\sigma}^2 \mid \mathbf{Z}^n)\} = \mathbb{E}(\widetilde{\sigma}^2),$$

where the second equality follows from the unbiasedness property.

We need the bound on $|\widehat{\sigma} - \widetilde{\sigma}|$ which can be expressed as a linear function of the absolute difference of the respective squares.

$$|\widehat{\sigma} - \widetilde{\sigma}| = |\widehat{\sigma} - \widetilde{\sigma}| \frac{|\widehat{\sigma} + \widetilde{\sigma}|}{|\widehat{\sigma} + \widetilde{\sigma}|} \leq \epsilon^{-1} |\widehat{\sigma}^2 - \widetilde{\sigma}^2|, \text{ if } \widehat{\sigma} + \widetilde{\sigma} > \epsilon > 0.$$

Hence, it is enough to show the bound on $|\widehat{\sigma}^2 - \widetilde{\sigma}^2|$. Equivalently, we can show that $\mathbb{E}|\widehat{\sigma}^2 - \widetilde{\sigma}^2|^2 \lesssim n^{-1}$. Evaluating this quantity by the law of iterated expectation

$$\begin{aligned}
\mathbb{E}\left(\widehat{\sigma}^2 - \widetilde{\sigma}^2\right)^2 &= \mathbb{E}\left[\mathbb{E}\left\{\left(\widehat{\sigma}^2 - \widetilde{\sigma}^2\right)^2 | \mathbf{Z}^n\right\}\right] \\
&= \mathbb{E}\left\{\mathbb{E}\left(\widehat{\sigma}^4 + \widetilde{\sigma}^4 - 2\widehat{\sigma}^2\widetilde{\sigma}^2 | \mathbf{Z}^n\right)\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left(\widehat{\sigma}^4 | \mathbf{Z}^n\right) - \widetilde{\sigma}^4\right\},
\end{aligned}$$

where the last equality follows from $\mathbb{E}(\widehat{\sigma}^2|\mathbf{Z}^n) = \widetilde{\sigma}^2$. Next, we evaluate $\widehat{\sigma}^4$.

$$
\begin{aligned}
\widehat{\sigma}^4 =& \left(\widehat{\sigma}^2\right)^2 \\
=& \frac{N^2}{(N-1)^2}\left\{\frac{\sum_i \widehat{\phi}^2}{N} - \left(\frac{\sum_i \widehat{\phi}}{N}\right)^2\right\}^2 \\
=& \frac{N^2}{(N-1)^2}\left\{\frac{\sum_i \widehat{\phi}_i^4 + \sum_{i\neq j}\widehat{\phi}_i^2\widehat{\phi}_j^2}{N^2} - 2\frac{\sum_i \widehat{\phi}_i^4 + \sum_{i\neq j}\widehat{\phi}_i^2\widehat{\phi}_j^2 + 2\sum_{i\neq j}\widehat{\phi}_i^3\widehat{\phi}_j + \sum_{i\neq j\neq k}\widehat{\phi}_i^2\widehat{\phi}_j\widehat{\phi}_k}{N^3}\right. \\
&\left. + \frac{\sum_i \widehat{\phi}_i^4 + 3\sum_{i\neq j}\widehat{\phi}_i^2\widehat{\phi}_j^2 + 4\sum_{i\neq j}\widehat{\phi}_i^3\widehat{\phi}_j + 6\sum_{i\neq j\neq k}\widehat{\phi}_i^2\widehat{\phi}_j\widehat{\phi}_k + \sum_{i\neq j\neq k\neq l}\widehat{\phi}_i\widehat{\phi}_j\widehat{\phi}_k\widehat{\phi}_l}{N^4}\right\}.
\end{aligned}
$$

Let $\mu_m = \mathbb{E}(\widehat{\phi}^m \mid \mathbf{Z}^n)$. Thus, $\mathbb{E}(\widehat{\sigma}^2 \mid \mathbf{Z}^n) = \widetilde{\sigma}^2 = \mu_2 - \mu_1^2$.

$$
\begin{aligned}
\mathbb{E}\left(\widehat{\sigma}^4 \mid \mathbf{Z}^n\right) =& \frac{N^2}{(N-1)^2}\left\{\frac{N\mu_4 + N(N-1)\mu_2^2}{N^2}\right. \\
& - 2\frac{N\mu_4 + N(N-1)\mu_2^2 + 2N(N-1)\mu_3\mu_1 + N(N-1)(N-2)\mu_2\mu_1^2}{N^3} \\
& + \frac{\mu_4 + 3(N-1)\mu_2^2 + 4(N-1)\mu_3\mu_1 + 6(N-1)(N-2)\mu_2\mu_1^2}{N^3} \\
& \left. + \frac{(N-1)(N-2)(N-3)\mu_1^4}{N^3}\right\} \\
=& \frac{\mu_4}{(N-1)^2}\left\{N - 2 + \frac{1}{N}\right\} + \frac{\mu_2^2}{N-1}\left\{N - 2 + \frac{3}{N}\right\} + \frac{\mu_3\mu_1}{N-1}\left\{-2 + \frac{4}{N}\right\} \\
& + \frac{\mu_2\mu_1^2}{N-1}\left\{-2(N-2) + \frac{6(N-2)}{N}\right\} + \frac{\mu_1^4}{N-1}\frac{(N-2)(N-3)}{N} \\
=& \frac{\mu_4}{N} + \frac{\mu_2^2(N^2-2N+3)}{N(N-1)} - 2\frac{\mu_3\mu_1(N-2)}{N(N-1)} - \frac{2\mu_2\mu_1^2(N-2)(N-3)}{N(N-1)} \\
& + \frac{\mu_1^4(N-2)(N-3)}{N(N-1)}.
\end{aligned}
$$

Thus, combining the results

$$
\begin{aligned}
& \mathbb{E}\left(\widehat{\sigma}^4 \middle| \mathbf{Z}^n\right) - \widetilde{\sigma}^4 \\
=& \frac{\mu_4}{N} + \frac{\mu_2^2(N^2-2N+3)}{N(N-1)} - 2\frac{\mu_3\mu_1(N-2)}{N(N-1)} - \frac{2\mu_2\mu_1^2(N-2)(N-3)}{N(N-1)} + \frac{\mu_1^4(N-2)(N-3)}{N(N-1)} \\
& - (\mu_2 - \mu_1^2)^2, \ \text{ since } \mathbb{E}(\widehat{\sigma}^2 \mid \mathbf{Z}^n) = \widetilde{\sigma}^2 = \mu_2 - \mu_1^2 \\
=& \frac{\mu_4}{N} - \frac{\mu_2^2(N-3)}{N(N-1)} + \frac{4\mu_2\mu_1^2(2N-3)}{N(N-1)} - \frac{2\mu_1^4(2N-3)}{N(N-1)} - 2\frac{\mu_3\mu_1(N-2)}{N(N-1)} \\
\lesssim& N^{-1} \lesssim n^{-1}.
\end{aligned}
$$

Thus, $\left|\widehat{\sigma}^2 - \widetilde{\sigma}^2\right| \lesssim \frac{1}{\sqrt{n}}$. □

Proof of Theorem 2.4. The estimate of population size, $\widehat{n} = N/\widehat{\psi}$ depends on two random quantities (i) the number of observations N and (ii) the estimate of the capture probability ψ.

Calculation of mean and variance of $\widehat{n}$

First, we re-write $\frac{N}{\psi}$ as a sample average for ease of calculation. As mentioned in the proof of Theorem 2.2, we use $\mathbb{Q}\widehat{\varphi}$ to denote the population average conditioned on training sample $\mathbf{Z}^n$ i.e., $\int \widehat{\varphi}(\mathbf{z})d\mathbb{Q}(\mathbf{z})$.

$$\begin{aligned}
\widehat{n} - n = \; & N/\widehat{\psi} - N/\psi + N/\psi - n \\
= \; & N\left\{(\mathbb{Q}_N - \mathbb{Q})\widehat{\varphi} + \widehat{R}_2\right\} + N/\psi - n, \text{ (follows from proof of Theorem 2.2)} \\
= \; & n\mathbb{P}_n\mathbb{1}(\mathbf{Y} \neq \mathbf{0})\left(\widehat{\varphi} - \mathbb{Q}\widehat{\varphi} + \widehat{R}_2 + \psi^{-1}\right) - n, \text{ since } N = n\mathbb{P}_n\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) \\
= \; & n\mathbb{P}_n \underbrace{\left[\mathbb{1}(\mathbf{Y} \neq \mathbf{0})\left(\widehat{\varphi} - \mathbb{Q}\widehat{\varphi}\right) + \{\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) - \psi\}\left(\widehat{R}_2 + \psi^{-1}\right)\right]}_{\zeta} + n\psi\widehat{R}_2.
\end{aligned}$$

Thus, $\widehat{n} - n - n\psi\widehat{R}_2$ is a sample average $n\mathbb{P}_n\zeta$. Moreover, when we condition on the training sample $\mathbf{Z}^n$, the ζ's are i.i.d. We present the conditional mean and variance of ζ below.

$$\mathbb{E}(\zeta|\mathbf{Z}^n) = \mathbb{E}\{\mathbb{E}(\zeta|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n)|\mathbf{Z}^n\} = 0.$$

$$\begin{aligned}
var(\zeta|\mathbf{Z}^n) &= var\{\mathbb{E}(\zeta|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n)\} + \mathbb{E}\{var(\zeta|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n)\} \\
&= var\left[\{\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) - \psi\}\left(\widehat{R}_2 + \psi^{-1}\right)|\mathbf{Z}^n\right] + \mathbb{E}\left[\mathbb{1}(\mathbf{Y} \neq \mathbf{0})var\left(\widehat{\varphi}|\mathbf{Z}^n\right)|\mathbf{Z}^n\right] \\
&= \psi(1-\psi)(\widehat{R}_2 + \psi^{-1})^2 + \psi \; var\left(\widehat{\varphi}|\mathbf{Z}^n\right) \\
&= \frac{1-\psi}{\psi}(\psi\widehat{R}_2 + 1)^2 + \psi\widetilde{\varsigma}^2, \quad \text{defining } \widetilde{\varsigma}^2 = var(\widehat{\varphi}|\mathbf{Z}^n).
\end{aligned}$$

The population expectation of $\widehat{n} - n$ is $n\psi\mathbb{E}(\widehat{R}_2)$. And the population variance is presented below.

$$\begin{aligned}
var(\widehat{n} - n) &= var\{\mathbb{E}(\widehat{n} - n|\mathbf{Z}^n)\} + \mathbb{E}\{var(\widehat{n} - n|\mathbf{Z}^n)\} \\
&= var\{\mathbb{E}(n\mathbb{P}_n\zeta + n\psi\widehat{R}_2|\mathbf{Z}^n)\} + \mathbb{E}\{var(n\mathbb{P}_n\zeta + n\psi\widehat{R}_2|\mathbf{Z}^n)\} \\
&= var(n\psi\widehat{R}_2) + \mathbb{E}\{n \; var(\zeta|\mathbf{Z}^n)\} \\
&= n^2\psi^2 var(\widehat{R}_2) + n\psi\mathbb{E}\left(\widetilde{\varsigma}^2\right) + n\frac{1-\psi}{\psi}\mathbb{E}(\psi\widehat{R}_2 + 1)^2.
\end{aligned}$$

Thus, $\mathbb{E}(\widehat{n} - n)^2 = n^2\psi^2\mathbb{E}(\widehat{R}_2^2) + n\psi\mathbb{E}\left(\widetilde{\varsigma}^2\right) + n\frac{1-\psi}{\psi}\mathbb{E}(\psi\widehat{R}_2 + 1)^2$.

If $\mathbb{E}|\widehat{R}_2|$ is sufficiently small, then $\widehat{n} - n$ has expectation approximately 0 and variance approximately $n\psi\varsigma^2 + n(1-\psi)/\psi$, where $\varsigma^2 = var(\varphi)$.

Approximate normality

We define the estimated variance of $\widehat{n} - n - n\psi\widehat{R}_2$ conditioned on the training sample as $\widehat{n}\widehat{\tau}^2$, where $\widehat{\tau}^2 = \widehat{\psi}\widehat{\varsigma}^2 + \frac{1-\widehat{\psi}}{\widehat{\psi}}$ and $\widehat{\varsigma}^2 = \widehat{var}(\widehat{\varphi}|\mathbf{Z}^n)$ is the unbiased estimator of $\widetilde{\varsigma}^2$ conditioned on the training data. Let $\widetilde{\tau}^2 = var(\zeta|\mathbf{Z}^n) = \psi\widetilde{\varsigma}^2 + \frac{1-\psi}{\psi}(\psi\widehat{R}_2 + 1)^2$.

We ultimately want to see the error in normal approximation for $\frac{\widehat{n}-n}{\widehat{\tau}\sqrt{\widehat{n}}}$.

By Berry-Esseen we have for any t' and n that

$$\Phi(t') - \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}} \le \mathbb{P}\left(\frac{\widehat{n} - n - n\psi\widehat{R}_2}{\widetilde{\tau}\sqrt{n}} \le t' \;\Big|\; \mathbf{Z}^n \right) \le \Phi(t') + \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}}$$

Taking $t' = \frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}}t - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}}$ and $\rho = \mathbb{E}\left(|\zeta|^3 \big| \mathbf{Z}^n\right)$, this implies

$$\Phi\left(\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}}t - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}} \right) - \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}} \le \mathbb{P}\left(\frac{\widehat{n}-n}{\widehat{\tau}\sqrt{\widehat{n}}} \le t \;\Big|\; \mathbf{Z}^n \right) \le \Phi\left(\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}}t - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}} \right) + \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}}$$

or equivalently

$$\begin{aligned} \Phi\left(t\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}} \right) - \Phi(t) - \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}} &\le \mathbb{P}\left(\frac{\widehat{n}-n}{\widehat{\tau}\sqrt{\widehat{n}}} \le t \;\Big|\; \mathbf{Z}^n \right) - \Phi(t) \\ &\le \Phi\left(t\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}} \right) - \Phi(t) + \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}}. \end{aligned}$$

Now note by the mean value theorem, for some t_n between t and $t\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}}$, we have that

$$\begin{aligned} \left| \Phi\left(t\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}} \right) - \Phi(t) \right| &= \left| \Phi'(t_n) \left\{ \left(\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - 1 \right) t - \frac{\psi\widehat{R}_2}{\widetilde{\tau}/\sqrt{n}} \right\} \right| \\ &\le \frac{1}{\sqrt{2\pi}} \left(\left| \frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - 1 \right| |t| + \frac{\psi|\widehat{R}_2|}{\widetilde{\tau}/\sqrt{n}} \right) \equiv \Delta_n \end{aligned}$$

where the second inequality used the fact that $\sup_t \Phi'(t) \le 1/\sqrt{2\pi}$ and the triangle inequality.

Therefore

$$-\Delta_n - \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}} \le \mathbb{P}\left(\frac{\widehat{n}-n}{\widehat{\tau}\sqrt{\widehat{n}}} \le t \;\Big|\; \mathbf{Z}^n \right) - \Phi(t) \le \Delta_n + \frac{C\rho}{\widetilde{\tau}^3\sqrt{n}}$$

This implies by iterated expectation that

$$\left| \mathbb{P}\left(\frac{\widehat{n}-n}{\widehat{\tau}\sqrt{\widehat{n}}} \le t \right) - \Phi(t) \right| \le \sqrt{\frac{1}{2\pi}} \left\{ \frac{\sqrt{n}\psi\mathbb{E}|\widehat{R}_2|}{\widetilde{\tau}} + |t|\mathbb{E}\left(\left| \frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - 1 \right| \right) \right\} + \frac{C}{\sqrt{n}}\mathbb{E}\left(\frac{\rho}{\widetilde{\tau}^3} \right).$$

Bound on $\mathbb{E}\left|\frac{\widehat{\tau}\sqrt{\widehat{\mathbf{n}}}}{\widetilde{\tau}\sqrt{\mathbf{n}}}-\mathbf{1}\right|$

It is easy to see that

$$\frac{1}{\sqrt{n}\widetilde{\tau}}\left|\widehat{\tau}\sqrt{\widehat{n}}-\widetilde{\tau}\sqrt{n}\right| = \frac{1}{\sqrt{n}\widetilde{\tau}}\left|\widehat{\tau}\sqrt{\widehat{n}}-\widetilde{\tau}\sqrt{n}\right| \frac{\left|\widehat{\tau}\sqrt{\widehat{n}}+\widetilde{\tau}\sqrt{n}\right|}{\left|\widehat{\tau}\sqrt{\widehat{n}}+\widetilde{\tau}\sqrt{n}\right|} \leq \frac{\left|\widehat{\tau}^2\widehat{n}-\widetilde{\tau}^2 n\right|}{n\widetilde{\tau}^2}.$$

By simple algebra, we can bound the quantity on the right hand side above as follows.

$$\begin{aligned}
\frac{1}{n\widetilde{\tau}^2}\left|\widehat{\tau}^2\frac{N}{\widehat{\psi}}-\widetilde{\tau}^2 n\right| &= \frac{1}{n\widetilde{\tau}^2}\left|\widehat{\psi}\widehat{\varsigma}^2\frac{N}{\widehat{\psi}}+\frac{1-\widehat{\psi}}{\widehat{\psi}}\frac{N}{\widehat{\psi}}-\psi\widetilde{\varsigma}^2 n-\frac{1-\psi}{\psi}(\psi\widehat{R}_2+1)^2 n\right| \\
&\leq \frac{N}{n\widetilde{\tau}^2}|\widehat{\varsigma}^2-\widetilde{\varsigma}^2|+\frac{1}{n\psi}|N-n\psi|+\frac{N}{n\widetilde{\tau}^2}\left|\frac{1}{\widehat{\psi}^2}-\frac{1}{\psi^2}-\frac{1}{\widehat{\psi}}+\frac{1}{\psi}\right| \\
&\quad +\frac{1}{\widetilde{\tau}^2}(1-\psi)\left(\psi\widehat{R}_2^2+2|\widehat{R}_2|\right).
\end{aligned}$$

The last inequality follows using the definition of $\widetilde{\tau}$ and triangle inequality. To evaluate the third term, we will use the relation $\widehat{\psi}^{-1}-\psi^{-1}=(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}+\widehat{R}_2$. Thus,

$$\begin{aligned}
&\frac{1}{\widehat{\psi}^2}-\frac{1}{\psi^2}-\frac{1}{\widehat{\psi}}+\frac{1}{\psi} \\
&= \left\{\psi^{-1}+(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}+\widehat{R}_2\right\}^2-\psi^{-2}-(\mathbb{Q}-\mathbb{Q}_N)\widehat{\varphi}-\widehat{R}_2 \\
&= \{(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}\}^2+\widehat{R}_2^2+(2\psi^{-1}+2\widehat{R}_2-1)(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}+(2\psi^{-1}-1)\widehat{R}_2.
\end{aligned}$$

Similar to the proof of Theorem 2.2,

$$\mathbb{E}\left\{\left|(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}\right|\middle|\mathbf{Z}^n\right\} \leq \left(\mathbb{E}\left[\left\{(\mathbb{Q}_N-\mathbb{Q})\,\widehat{\varphi}\right\}^2\middle|\mathbf{Z}^n\right]\right)^{1/2} = \frac{\widetilde{\varsigma}}{\sqrt{N}} \leq \frac{\widetilde{\tau}}{\sqrt{\psi N}}.$$

The last bound follows by the relation between $\widetilde{\tau}$ and $\widetilde{\varsigma}$. Thus,

$$\begin{aligned}
&\mathbb{E}\left\{\left|\frac{1}{\widehat{\psi}^2}-\frac{1}{\psi^2}-\frac{1}{\widehat{\psi}}+\frac{1}{\psi}\right|\middle|\mathbf{Z}^n\right\} \\
&\leq \mathbb{E}\left[\{(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}\}^2\middle|\mathbf{Z}^n\right]+\widehat{R}_2^2+\left|2\psi^{-1}+2\widehat{R}_2-1\right|\mathbb{E}\left\{\left|(\mathbb{Q}_N-\mathbb{Q})\widehat{\varphi}\right|\middle|\mathbf{Z}^n\right\}+(2\psi^{-1}-1)\left|\widehat{R}_2\right| \\
&\leq \frac{\widetilde{\tau}^2}{N\psi}+\widehat{R}_2^2+\left|2\psi^{-1}+2\widehat{R}_2-1\right|\frac{\widetilde{\tau}}{\sqrt{N\psi}}+(2\psi^{-1}-1)\left|\widehat{R}_2\right|.
\end{aligned}$$

Combining the results above, we get the following bound.

$$\begin{aligned}\mathbb{E}\left|\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}}-1\right| &\le \mathbb{E}\left(\frac{N}{n\widetilde{\tau}^2}|\widehat{\varsigma}^2-\widetilde{\varsigma}^2|\right)+\frac{1}{n\psi}\mathbb{E}|N-n\psi|+\frac{1}{n\psi}+\mathbb{E}\left(\frac{N\widehat{R}_2^2}{n\widetilde{\tau}^2}\right)\\ &+\mathbb{E}\left(\left|2\psi^{-1}+2\widehat{R}_2-1\right|\frac{\sqrt{N}}{n\sqrt{\psi}\widetilde{\tau}}\right)+(2\psi^{-1}-1)\mathbb{E}\left(\frac{N|\widehat{R}_2|}{n\widetilde{\tau}^2}\right)+(1-\psi)\mathbb{E}\left(\frac{\psi\widehat{R}_2^2+2|\widehat{R}_2|}{\widetilde{\tau}^2}\right).\end{aligned}$$

Next, using the inequality that $N \le n$ and $\mathbb{E}|N-n\psi| \le \left\{\mathbb{E}\left(N-n\psi\right)^2\right\}^{1/2} = \{n\psi(1-\psi)\}^{1/2}$, we get

$$\begin{aligned}\mathbb{E}\left|\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}}-1\right| &\le \mathbb{E}\left(\frac{|\widehat{\varsigma}^2-\widetilde{\varsigma}^2|}{\widetilde{\tau}^2}\right)+\frac{\sqrt{1-\psi}}{\sqrt{n\psi}}+\frac{1}{n\psi}+\mathbb{E}\left(\frac{\widehat{R}_2^2}{\widetilde{\tau}^2}\right)+\mathbb{E}\left\{\frac{\left(2\psi^{-1}-1+2|\widehat{R}_2|\right)}{\sqrt{n\psi}\widetilde{\tau}}\right\}\\ &\quad+(2\psi^{-1}-1)\mathbb{E}\left(\frac{|\widehat{R}_2|}{\widetilde{\tau}^2}\right)+(1-\psi)\mathbb{E}\left(\frac{\psi\widehat{R}_2^2+2|\widehat{R}_2|}{\widetilde{\tau}^2}\right)\\ &= \mathbb{E}\left(\frac{|\widehat{\varsigma}^2-\widetilde{\varsigma}^2|}{\widetilde{\tau}^2}\right)+\frac{\sqrt{1-\psi}}{\sqrt{n\psi}}+\frac{1}{n\psi}+\{\psi\left(1-\psi\right)+1\}\,\mathbb{E}\left(\frac{\widehat{R}_2^2}{\widetilde{\tau}^2}\right)+\mathbb{E}\left(\frac{2\psi^{-1}-1}{\sqrt{n\psi}\widetilde{\tau}}\right)\\ &\quad+\mathbb{E}\left\{\frac{|\widehat{R}_2|}{\widetilde{\tau}^2}\left(\frac{2\widetilde{\tau}}{\sqrt{n\psi}}+2\psi^{-1}+1-2\psi\right)\right\}\\ &\le \mathbb{E}\left(\frac{|\widehat{\varsigma}^2-\widetilde{\varsigma}^2|}{\widetilde{\tau}^2}\right)+\frac{\sqrt{1-\psi}}{\sqrt{n\psi}}+\frac{1}{n\psi}+\mathbb{E}\left(\frac{2\psi^{-3/2}}{\sqrt{n}\widetilde{\tau}}\right)+2\mathbb{E}\left(\frac{\widehat{R}_2^2}{\widetilde{\tau}^2}\right)\\ &\quad+\mathbb{E}\left\{\frac{|\widehat{R}_2|}{\widetilde{\tau}^2}\left(\frac{2\widetilde{\tau}}{\sqrt{n\psi}}+2\psi^{-1}+1-2\psi\right)\right\}.\end{aligned}$$

Next, we obtain the asymptotic bound on the absolute difference in the cumulative functions.

Let $\mathbb{E}|\widehat{R}_2| \lesssim n^{-2\beta}$. Following the proof of Theorem 2.3 we have $\mathbb{E}|\widehat{\varsigma}^2-\widetilde{\varsigma}^2| \lesssim n^{-1/2}$. Thus, if $\widetilde{\tau} \gtrsim 1$ with probability 1 and $\psi \ge \epsilon > 0$ then

$$\mathbb{E}\left|\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}}-1\right| \lesssim n^{-1/2}+n^{-2\beta}.$$

Further, if $\mathbb{E}\left(\frac{\rho}{\widetilde{\tau}^3}\right) < c$ for some finite constant c,

$$\left|\mathbb{P}\left(\frac{\widehat{n}-n}{\widehat{\tau}\sqrt{\widehat{n}}} \le t\right)-\Phi(t)\right| \lesssim n^{(1-4\beta)/2}+n^{-1/2}.$$

Coverage error

The $(1-\alpha)\%$ estimated CI for n is $\widehat{n} \pm z_{\alpha/2}\widehat{\tau}\sqrt{\widehat{n}}$.

$$\begin{aligned}
&\left|\mathbb{P}\left(\widehat{n} - z_{\alpha/2}\widehat{\tau}\sqrt{\widehat{n}} \leq n \leq \widehat{n} + z_{\alpha/2}\widehat{\tau}\sqrt{\widehat{n}}\right) - (1-\alpha)\right| \\
=& \left|\mathbb{P}\left(\frac{n-\widehat{n}}{\widehat{\tau}\sqrt{\widehat{n}}} \leq z_{\alpha/2}\right) - \mathbb{P}\left(\frac{n-\widehat{n}}{\widehat{\tau}\sqrt{\widehat{n}}} \leq -z_{\alpha/2}\right) - \Phi(z_{\alpha/2}) + \Phi(-z_{\alpha/2})\right| \\
\leq& \left|\mathbb{P}\left(\frac{n-\widehat{n}}{\widehat{\tau}\sqrt{\widehat{n}}} \leq z_{\alpha/2}\right) - \Phi(z_{\alpha/2})\right| + \left|\mathbb{P}\left(\frac{n-\widehat{n}}{\widehat{\tau}\sqrt{\widehat{n}}} \leq -z_{\alpha/2}\right) - \Phi(-z_{\alpha/2})\right| \\
\leq& \sqrt{\frac{2}{\pi}}\left\{\frac{\sqrt{n}\psi\mathbb{E}|\widehat{R}_2|}{\widetilde{\tau}} + |z_{\alpha/2}|\mathbb{E}\left(\left|\frac{\widehat{\tau}\sqrt{\widehat{n}}}{\widetilde{\tau}\sqrt{n}} - 1\right|\right)\right\} + \frac{2C}{\sqrt{n}}\mathbb{E}\left(\frac{\rho}{\widetilde{\tau}^3}\right) \\
\lesssim& \, n^{(1-4\beta)/2} + n^{-1/2}.
\end{aligned}$$

□

A.1.1 Two lists vs multiple lists

The data-set under consideration has K lists. The proposed method focuses on the conditional independence assumption of two lists ($Y_1 \perp\!\!\!\perp Y_2 \mid \mathbf{X}$). The question is, when there are more than two lists, whether one should ignore the other $K-2$ lists (i.e. delete all rows that appear in neither list 1 nor list 2, but only in one or more of the remaining lists), or keep them. To answer this question, we evaluate the variance under these two cases. Below we present the variance of the estimated population size when ψ is known.

$$var(\widehat{n}) = var\left(\frac{N}{\psi}\right) = \frac{n\psi(1-\psi)}{\psi^2} = n\left(\frac{1}{\psi} - 1\right).$$

1. **Only two lists used**
   $\psi = \mathbb{P}(Y_1 \neq 0 \text{ or } Y_2 \neq 0)$.
   $\gamma(\mathbf{x}) = \mathbb{P}((Y_1, Y_2) \neq (0,0) \mid \mathbf{X} = \mathbf{x})$.

2. **All lists used**
   $\psi = \mathbb{P}(Y_1 \neq 0 \text{ or } Y_2 \neq 0 \text{ or } \ldots Y_K \neq 0)$.
   $\gamma(\mathbf{x}) = \mathbb{P}(\mathbf{Y} \neq \mathbf{0} \mid \mathbf{X} = \mathbf{x})$.
   The ψ and the $\gamma(\mathbf{x})$ for this case are larger than the ones for the two list case above.

It is easy to see that $var(\widehat{n})$ is smaller when all K lists are used since we observe more individuals.

A.2 TMLE

In this section we present the targeted maximum likelihood algorithm to estimate the capture probability ψ. We iteratively obtain estimate for the nuisance functions i.e., the q-probabilities. These estimates, $\widehat{q}^*$'s are used to obtain $\widehat{\psi}_{tmle}$.

Algorithm 1: TMLE algorithm for estimating ψ

```
1. Obtain initial estimates of $q_{12}(\mathbf{x})$, $q_1(\mathbf{x})$ and $q_2(\mathbf{x})$, denoted $\widehat{q}_{12,0}(\mathbf{x})$, $\widehat{q}_{1,0}(\mathbf{x})$ and $\widehat{q}_{2,0}(\mathbf{x})$. Set $t = 0$.
2. At step $t$, construct clever covariates:
    (a) $\mathbf{H}_{12,t} = \frac{\widehat{q}_{1,t}(\mathbf{X})\widehat{q}_{2,t}(\mathbf{X})}{\widehat{q}_{12,t}(\mathbf{X})^2} - \frac{\widehat{q}_{1,t}(\mathbf{X})}{\widehat{q}_{12,t}(\mathbf{X})} - \frac{\widehat{q}_{2,t}(\mathbf{X})}{\widehat{q}_{12,t}(\mathbf{X})}$
    (b) $\mathbf{H}_{1,t} = \frac{\widehat{q}_{2,t}(\mathbf{X})}{\widehat{q}_{12,t}(\mathbf{X})}$
    (c) $\mathbf{H}_{2,t} = \frac{\widehat{q}_{1,t}(\mathbf{X})}{\widehat{q}_{12,t}(\mathbf{X})}$.
3. Regress $Y_1Y_2$ on $\mathbf{H}_{12,t}$ using a no-intercept logistic model with $\text{logit}\{\widehat{q}_{12,t}(\mathbf{X})\}$ as offset, obtaining estimated coefficient $\widehat{\beta}_{12,t}$. Set $\widehat{q}_{12,t+1}(\mathbf{X}) = \text{expit}\Big[\text{logit}\{\widehat{q}_{12,t}(\mathbf{X})\} + \widehat{\beta}_{12}\mathbf{H}_{12,t}\Big]$.
4. Regress $Y_1(1 - Y_2)$ on $\mathbf{H}_{1,t}$ using a no-intercept logistic model with $\text{logit}\{\widehat{q}_{1,t}(\mathbf{X}) - \widehat{q}_{12,t+1}(\mathbf{X})\}$ as offset, obtaining estimated coefficient $\widehat{\beta}_{1,t}$. Set
    $\widehat{q}_{1,t+1}(\mathbf{X}) = min\Big\{\widehat{q}_{12,t+1}(\mathbf{X}) + \text{expit}\Big[\text{logit}\{\widehat{q}_{1,t}(\mathbf{X}) - \widehat{q}_{12,t+1}(\mathbf{X})\} + \widehat{\beta}_{1,t}\mathbf{H}_{1,t}\Big], 1 - \widehat{q}_{12,t+1}(\mathbf{X})\Big\}.$
5. Regress $Y_2(1 - Y_1)$ on $\mathbf{H}_{2,t}$ using a no-intercept logistic model with $\text{logit}\{\widehat{q}_{2,t}(\mathbf{x}) - \widehat{q}_{12,t+1}(\mathbf{x})\}$ as offset, obtaining estimated coefficient $\widehat{\beta}_{2,t}$. Set
    $\widehat{q}_{2,t+1}(\mathbf{X}) = min\Big\{\widehat{q}_{12,t+1}(\mathbf{X}) + \text{expit}\Big[\text{logit}\{\widehat{q}_{2,t}(\mathbf{X}) - \widehat{q}_{12,t+1}(\mathbf{X})\} + \widehat{\beta}_{2,t}\mathbf{H}_{2,t}\Big], 1 + \widehat{q}_{12,t+1}(\mathbf{X}) - \widehat{q}_{1,t+1}(\mathbf{X})\Big\}.$
6. Update $t \longrightarrow t + 1$. Repeat Steps 2 to 6 until convergence (e.g., until $\max_j |\widehat{\beta}_{j,t+1}| \leq \epsilon$).
Finally, set $\widehat{\psi}_{tmle} = \Big[\mathbb{Q}_N\Big\{\frac{\widehat{q}_1^*(\mathbf{X})\widehat{q}_2^*(\mathbf{X})}{\widehat{q}_{12}^*(\mathbf{X})}\Big\}\Big]^{-1}$, with $\widehat{q}_j^*$ estimates obtained after convergence.
```

Remark 27. *Step 5 in the algorithm can be modified so that for* $K = 2$, q_2 *is evaluated by* $\widehat{q}_{2,t+1}(\mathbf{x}) = 1 + \widehat{q}_{12,t+1} - \widehat{q}_{1,t+1}(\mathbf{x})$. *This step uses the relation that for* $K = 2$, $q_1 + q_2 - q_{12} = 1$.

A.3 Simulated data for varying total population size

The plug-in, the doubly robust and the TMLE are applied on the simulated data from section 2.6.1 total population size varying from 5000 to 25000. We focus on the case $\alpha = 0.25$ since it is the non-parametric convergence rate. The plots are presented in figure A.1. For each combination of (ψ, n), we simulated a dataset 500 times. The bias and RMSE of all three estimators decrease with the true total population size n and hence with the sample size N. However, the doubly robust and the targeted maximum likelihood estimators have smaller bias and RMSE. Similarly, the coverage of the total population size estimate for the plug-in estimator is much lower than the nominal coverage of 0.95 compared to the proposed methods' coverage.

A.4 Peru Internal Conflict Data 1980-2000

The data is collected by the Truth and Reconciliation Commission of Peru (Ball et al., 2003). It was further expanded by Rendon (2019a) with the addition of geographical parameters (Rendon, 2019b). The original dataset contains the geographic location in the form of Peru's UBIGEO codes. Rendon (2019a) added the continuous geographical coordinates for each region.

Ball et al. (2003) divided Peru into 58 stratas using the UBIGEO codes. We used the approach of Rendon (2019a) and calculated the latitude, longitude and area for a region (department or strata) using shape files from PERÚ (2014). For the latitude and longitude of a region, we averaged the latitude and longitude of the border of that region as in Rendon (2019a).

The victims that have no assigned strata are discarded by Ball et al. (2003). We present the statistics of these victims in table A.1.

Department	State	PCP-Shining Path	Others	Unidentified
No department available	396	137	11	369
Ayacucho	1257	21	102	38
Huancavelica	79	0	2	0
Junin	73	1	0	10
Lima	56	3	0	4
San Martin	76	0	3	1
Total	1937	162	118	422

Table A.1: This table shows the count of the victims with missing strata information by perpetrator and department. 75% of these victims have been captured by lists DP and ODH. 73% of the victims belong to the State.

List of covariates used to model the nuisance functions are as follows:

- age: numeric variable and takes 0 for missing age.

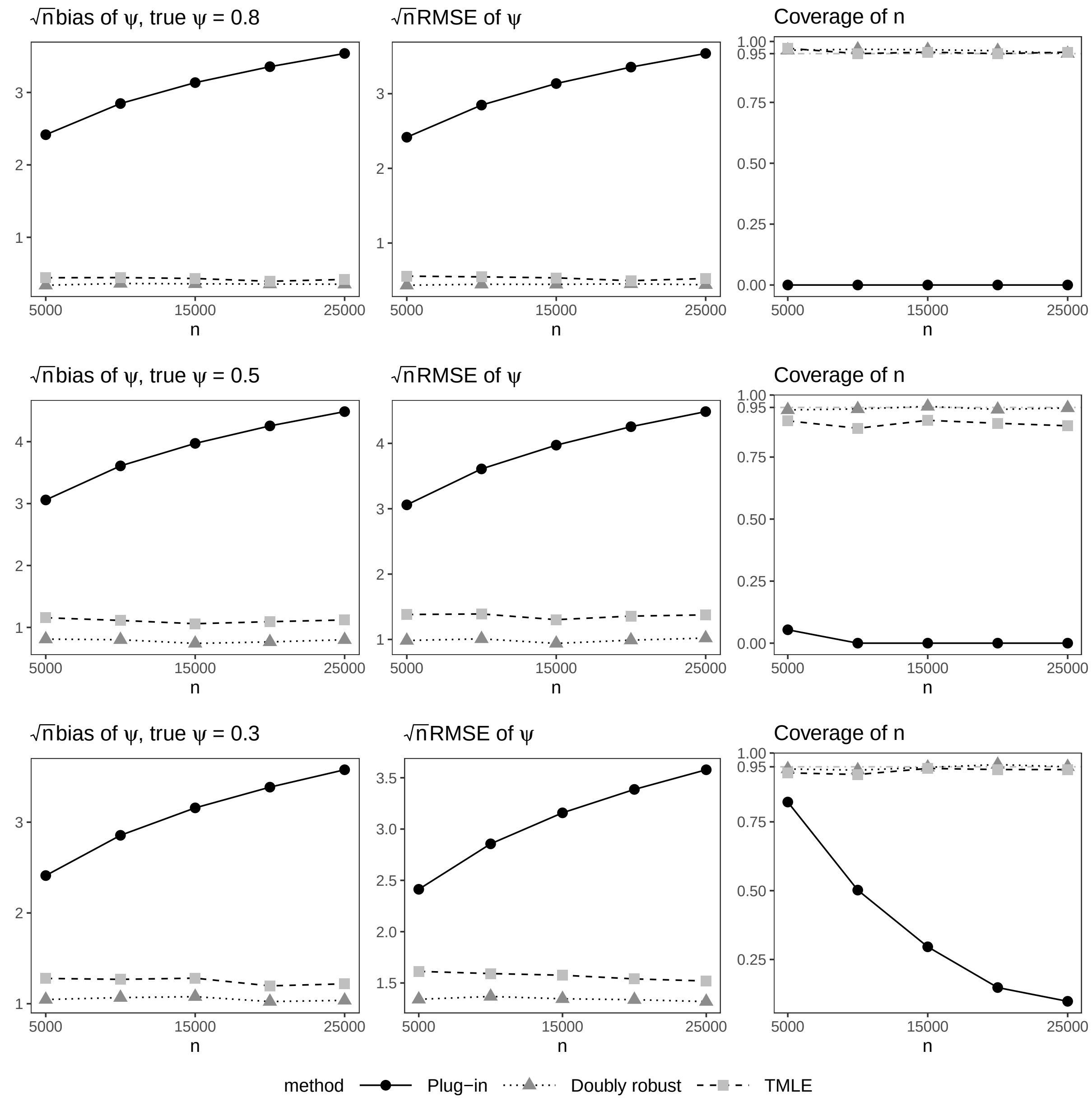

Figure A.1: Estimated bias, RMSE, and population size coverage, for simulated data with population size $n \in \{5000, 10000, 15000, 20000, 25000\}$, across true capture probability $\psi \in \{0.8, 0.5, 0.3\}$, and q-probability error rate $n^{-0.25}$.

- indicator for non-missing age: takes value 1 if age is present and 0 otherwise.
- gender: has levels male, female and others (including missing).
- situation: whether the victim was killed or disappeared.
- perpetrator: four levels indicating whether the individual is a victim of the State, Shining Path, others, or unidentified groups.
- indicator for non-missing department information: takes value 1 if department information is available for the individual and 0 otherwise.
- department latitude, longitude and area in hectares. For the individuals with missing department code, we use the average latitude, average longitude and median area of all the departments.
- strata code: 59 possible levels. Details and construction of the 58 strata are available in Ball et al. (2003)). Those with missing strata take value 59.
- strata latitude, longitude and area in hectares. For the individuals with missing strata code, we use the average latitude, average longitude and median area of all the strata.
- indicator of non-missing strata code.

A.4.1 Results

We present the exact estimated number of victims using our proposed doubly robust estimation in Table A.2. The difference in the estimated number of victims of the State and the Shining Path for the 25 departments and the seven geographic regions (Ball et al., 2003, see) are presented in Figure A.2. The State has a significantly higher estimated number of victims in department Ayacucho and the Northern region. The Shining Path has higher estimated number of victims in departments Junin and Puno.

Perpetrator	N	$\widehat{n}$	95% CI
State	11564	20756.00	[13775, 27737]
PCP-Shining Path	9243	13313.00	[10333, 16293]
Unidentified	3399	25749.00	[13384, 38114]
Total	24692	68874.00	[58543, 79204]

Table A.2: Observed and estimated numbers of killings and disappearances by perpetrator, using the proposed doubly robust method, with 95% confidence intervals.

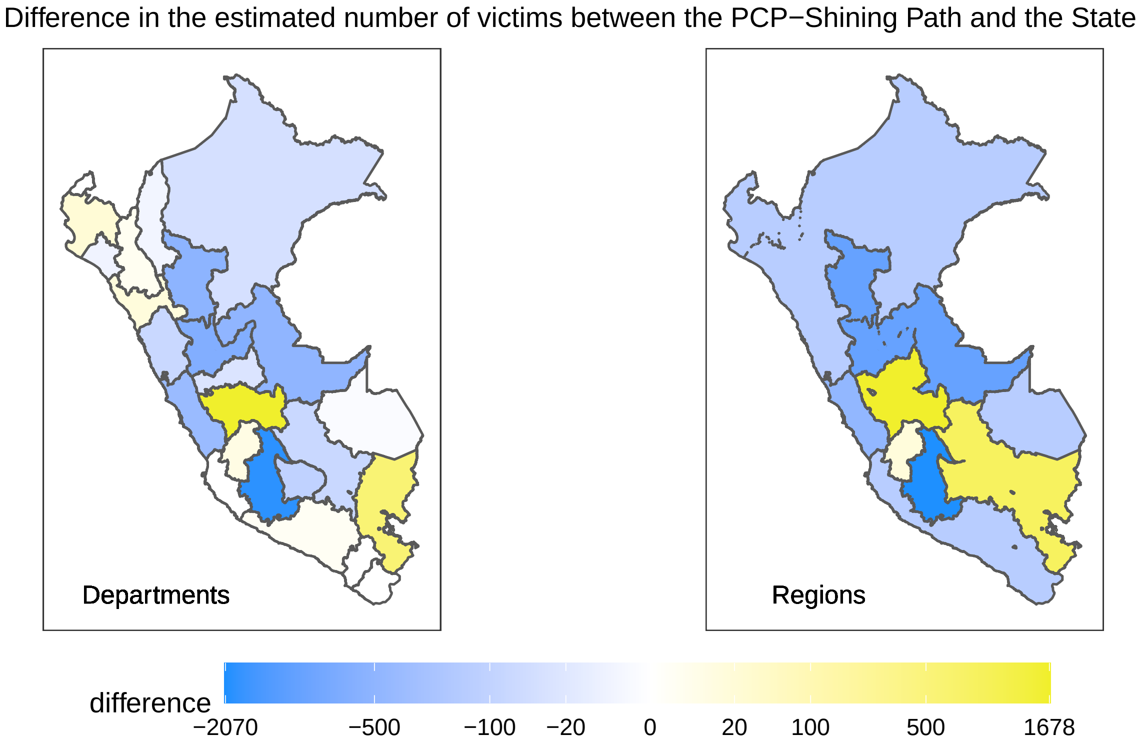
Difference in the estimated number of victims between the PCP−Shining Path and the State
Departments
Regions
difference
−2070
−500
−100
−20
0
20
100
500
1678

Figure A.2: Difference in the estimated number of killings by the PCP-Shining Path and the State in the 25 departments and the seven regions of Peru (Ball et al., 2003) from 1980-2000. The departments with comparatively higher number of victims for the State are in a darker shade and the ones with higher number of killings for the Shining Path are in a lighter shade.

Appendix B

Appendix for chapter 3

Proof of Theorem 4.1. We present the finite-sample error in the coverage guarantee that the proposed confidence interval contains the lower limit of ψ. The proof for ψ_u follows similarly.

We ultimately want to see the error in the coverage of the following Imbens and Manski (2004) $(1-\alpha) \times 100\%$ confidence interval

$$\left[\widehat{\psi}_l - \bar{C}_N \widehat{\sigma}_l/\sqrt{N},\ \widehat{\psi}_u + \bar{C}_N \widehat{\sigma}_u/\sqrt{N}\right].$$

Below we evaluate the finite sample coverage error of this estimated confidence interval. The error in coverage is

$$\left|(1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N \widehat{\sigma}_l/\sqrt{N} \le \psi \le \widehat{\psi}_u + \bar{C}_N \widehat{\sigma}_u/\sqrt{N}\right)\right|.$$

The quantity $\bar{C}_N$ is a stochastic quantity that depends on $\widehat{\psi}_l$, $\widehat{\psi}_u$, $\widehat{\sigma}_l$ and $\widehat{\sigma}_u$. Hence, to apply Berry-Esseen bound, we use a non-stochastic approximation of $\bar{C}_N$, which is c_N. c_N is a constant for a given n (also N) and training data $\mathbf{Z}^n$ unlike $\bar{C}_N$, and satisfies

$$\Phi\left(c_N + \sqrt{N}\frac{\psi_u - \psi_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi(-c_N) = 1-\alpha.$$

We also define an intermediate quantity $\widetilde{C}_N$ which satisfies

$$\Phi\left(\widetilde{C}_N + \sqrt{N}\frac{\widehat{\psi}_u - \widehat{\psi}_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi(-\widetilde{C}_N) = 1-\alpha.$$

Remark 28. *In the context of this paper, the difference $\psi_u - \psi_l$ is zero implies that the $\widetilde{\sigma}_u = \widetilde{\sigma}_l$ and also $\widehat{\psi}_u = \widehat{\psi}_l$. Thus, when the difference is zero, $\bar{C}_N = \widetilde{C}_N = c_N = z_{1-\alpha/2}$.*

Next, following the approach of Imbens and Manski (2004), we show that the estimated confidence interval contains ψ_u with probability $1-\alpha$ and some additional error. The proof for ψ_l follows similarly. Since, ψ

lies between ψ_u and ψ_l, the estimated interval will contain ψ with probability $1-\alpha$ and some additional error which is the maximum of the errors of ψ_u and ψ_l.

Below we evaluate the error is coverage of ψ_u by the estimated confidence interval.

$$\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N}\right) \\
&= (1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - c_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + c_N\widehat{\sigma}_u/\sqrt{N}\right) \\
&+ \mathbb{P}\left(\widehat{\psi}_l - c_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + c_N\widehat{\sigma}_u/\sqrt{N}\right) - \mathbb{P}\left(\widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N}\right) \\
&+ \mathbb{P}\left(\widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N}\right) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N}\right).
\end{aligned}$$

We need to show that following:

1. the first difference can be bounded above in the absolute sense, and

2. the second and the third differences are either positive or close to zero with probability close to 1.

Proof for the first difference

For the first difference

$$\begin{aligned}
&\left|\mathbb{P}\left(\widehat{\psi}_l - c_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + c_N\widehat{\sigma}_u/\sqrt{N}\right) - (1-\alpha)\right| \\
\le &\left|\mathbb{P}\left(\frac{\widehat{\psi}_l - \psi_l}{\widehat{\sigma}_l/\sqrt{N}} \le \frac{\psi_u - \psi_l}{\widehat{\sigma}_l/\sqrt{N}} + c_N\right) - \mathbb{P}\left(\frac{\widehat{\psi}_u - \psi_u}{\widehat{\sigma}_u/\sqrt{N}} \le -c_N\right) - (1-\alpha)\right| \\
\le &\left|\mathbb{P}\left(\frac{\widehat{\psi}_l - \psi_l}{\widehat{\sigma}_l/\sqrt{N}} \le c_N + \frac{\psi_u - \psi_l}{\widehat{\sigma}_l/\sqrt{N}}\right) - \Phi\left(c_N + \frac{\psi_u - \psi_l}{\widehat{\sigma}_l/\sqrt{N}}\right)\right| \\
+ &\left|\mathbb{P}\left(\frac{\widehat{\psi}_u - \psi_u}{\widehat{\sigma}_u/\sqrt{N}} \le -c_N\right) - \Phi\left(-c_N\right)\right| \\
+ &\left|\Phi\left(c_N + \frac{\psi_u - \psi_l}{\widehat{\sigma}_l/\sqrt{N}}\right) - \Phi\left(c_N + \sqrt{N}\frac{\psi_u - \psi_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right)\right| \\
&\text{since } \Phi\left(c_N + \sqrt{N}\frac{\psi_u - \psi_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi(-c_N) = 1-\alpha.
\end{aligned}$$

The first two terms are bounded above by Berry-Esseen and for the third term, we can use mean value theorem as follows,

$$\left|\Phi\left(c_N + \frac{\psi_u - \psi_l}{\widehat{\sigma}_l/\sqrt{N}}\right) - \Phi\left(c_N + \sqrt{N}\frac{\psi_u - \psi_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right)\right| = (\psi_u - \psi_l)\phi(t_3)\sqrt{N}\left|\frac{1}{\widehat{\sigma}_l} - \frac{1}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right|,$$

for some t_3 between $c_N + \frac{\psi_u - \psi_l}{\widehat{\sigma}_l/\sqrt{N}}$ and $c_N + \sqrt{N}\frac{\psi_u - \psi_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}$.

Since $c_N > 0$,

$$\phi(t_3) \le \phi\left(\sqrt{N}\frac{\psi_u - \psi_l}{\widehat{\sigma}_l \vee \widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) = \frac{1}{\sqrt{2\pi}} exp\left\{-N\frac{(\psi_u - \psi_l)^2}{2(\widehat{\sigma}_l \vee \widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}.$$

By Berry-Esseen we have for any t' and N that

$$\Phi(t') - \frac{C\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}} \le \mathbb{P}\left(\frac{\widehat{\psi}_l - \psi_l - \widehat{R}_{2,l}}{\widetilde{\sigma}_l/\sqrt{N}} \le t' \;\Big|\; \mathbf{Z}^n\right) \le \Phi(t') + \frac{C\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}$$

where $\rho_l = \mathbb{E}(|W_i| \mid \mathbf{Z}^n)^3$ when $\widehat{\psi}_l - \psi_l = \mathbb{Q}_N W_i$.
Taking $t' = \frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l}t - \frac{\sqrt{N}\widehat{R}_{2,l}}{\widetilde{\sigma}_l}$ and $\rho_l = \mathbb{E}\left(|\widehat{\psi}_l - \psi_l - \widehat{R}_{2,l}|^3\big|\mathbf{Z}^n\right)$, this implies

$$\Phi\left(\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l}t - \frac{\sqrt{N}\widehat{R}_{2,l}}{\widetilde{\sigma}_l}\right) - \frac{C\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}} \le \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\widehat{\sigma}_l} \le t \;\Big|\; \mathbf{Z}^n\right) \le \Phi\left(\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l}t - \frac{\sqrt{N}\widehat{R}_{2,l}}{\widetilde{\sigma}_l}\right) + \frac{C\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}.$$

Therefore, by the mean value theorem and the fact that $\psi(z) \le \frac{1}{\sqrt{2\pi}} \,\forall\, w$

$$\begin{aligned}\left|\mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\widehat{\sigma}_l} \le t \;\Big|\; \mathbf{Z}^n\right) - \Phi(t)\right| &\le \left|\Phi\left(t\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l} - \frac{\sqrt{N}\widehat{R}_{2,l}}{\widetilde{\sigma}_l}\right) - \Phi(t)\right| + \left|\frac{C\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}\right| \\ &\le \left|\frac{1}{\sqrt{2\pi}}\left(\left|\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l} - 1\right||t| + \frac{n\psi|\widehat{R}_2|}{\widetilde{\sigma}_l}\right)\right| + \left|\frac{C\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}\right|.\end{aligned}$$

This implies by iterated expectation that

$$\left|\mathbb{P}\left(\frac{\widehat{\psi}_l - \psi_l}{\widehat{\sigma}_l/\sqrt{n}} \le t\right) - \Phi(t)\right| \le \sqrt{\frac{1}{2\pi}}\left\{\frac{\mathbb{E}\left(\sqrt{N}|\widehat{R}_{2,l}|\right)}{\widetilde{\sigma}_l} + |t|\mathbb{E}\left(\left|\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l} - 1\right|\right)\right\} + C\mathbb{E}\left(\frac{\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}\right).$$

A similar result follows for $\widehat{\psi}_u$.

Proof for the second bound

The second difference can be re-written as

$$\begin{aligned}&\mathbb{P}(\widetilde{C}_N < c_N)\mathbb{P}\left(\widehat{\psi}_l - c_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \mid \widetilde{C}_N < c_N\right) \\ &+ \mathbb{P}(\widetilde{C}_N < c_N)\mathbb{P}\left(\widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + c_N\widehat{\sigma}_u/\sqrt{N} \mid \widetilde{C}_N < c_N\right) \\ &- \mathbb{P}(\widetilde{C}_N > c_N)\mathbb{P}\left(\widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_l - c_N\widehat{\sigma}_l/\sqrt{N} \mid \widetilde{C}_N > c_N\right) \\ &- \mathbb{P}(\widetilde{C}_N > c_N)\mathbb{P}\left(\widehat{\psi}_u + c_N\widehat{\sigma}_u/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N} \mid \widetilde{C}_N > c_N\right).\end{aligned}$$

The probabilities are bounded above by 1. Assuming that $\widehat{\psi}$ and $\widehat{\sigma}$ have continuous densities (to avoid high mass in a small area), it is sufficient to show that the positive terms are not too positive. We will show the following

$$\mathbb{P}(c_N - \widetilde{C}_N > \eta)$$

is bounded above. Define $\Delta = \psi_u - \psi_l$ and $\widehat{\Delta} = \widehat{\psi}_u - \widehat{\psi}_l$.

To prove the same, we will show that the following

$$\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right)$$

is bounded for a given $\eta > 0$. This however requires as additional assumption. We use the following finite sample assumption modifying the assumption in Imbens and Manski (2004) (assumption 1 (iii)).

Assumption 7. *For a given $\epsilon > 0$ and a constant c, there exists N_0 and $\upsilon > 0$ such that for all $N > N_0$*

$$\mathbb{P}\left(N^{\upsilon}|\widehat{\Delta} - \Delta| > c\right) < \epsilon.$$

Notice that we can break the event above into the following three terms.

$$\begin{aligned}
&\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta \\
&= \mathbb{1}(\widehat{\Delta} \leq \Delta) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\} \\
&+ \mathbb{1}(\widehat{\Delta} > \Delta, |\widehat{\Delta} - \Delta| \leq cN^{-\upsilon}) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\} \\
&+ \mathbb{1}(\widehat{\Delta} > \Delta, |\widehat{\Delta} - \Delta| > cN^{-\upsilon}) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\}.
\end{aligned}$$

The goal is to show that the probability of this event is not too large. So we start by maximizing this probability and show that it is bounded. The first term has zero probability and hence, can be dropped. For the second term, notice the following

$$\begin{aligned}
&\mathbb{1}\left\{\widehat{\Delta} > \Delta, |\widehat{\Delta} - \Delta| \leq cN^{-\upsilon}, \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\} \\
&\leq \mathbb{1}\left\{|\widehat{\Delta} - \Delta| \leq cN^{-\upsilon}, \phi\left(\frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) \times \frac{\sqrt{N}|\widehat{\Delta} - \Delta|}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l} > \eta\right\},
\end{aligned}$$

where the second bound follows by mean value theorem and the properties of the normal density ϕ. Now, by Markov's inequality we get the following bound

$$\frac{1}{\eta} \times \phi\left(\frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) \times \frac{cN^{\frac{1}{2}-\upsilon}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l} = \frac{cN^{\frac{1}{2}-\upsilon}}{\eta\sqrt{2\pi}(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\ exp\left\{-\frac{N\Delta^2}{2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}.$$

For the third term, it is easy to see that it is bounded above by

$$\mathbb{P}(|\widehat{\Delta} - \Delta| > cN^{-\upsilon}) < \frac{1}{\sqrt{N}},$$

when N is sufficiently large and c is a constant chosen appropriately.

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right) \\
&\leq \frac{cN^{\frac{1}{2}-\upsilon}}{\eta\sqrt{2\pi}(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\ exp\left\{-\frac{N\Delta^2}{2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\} + \frac{1}{\sqrt{N}}.
\end{aligned}$$

Next to show that $P(c_N - \widetilde{C}_N > \eta)$ is bounded above, we use mean value theorem. We will use the previously proved result and show that it is equivalent to $\mathbb{P}(c_N - \widetilde{C}_N > \eta)$.

$$\begin{aligned}
&\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) \\
&= \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi(-\widetilde{C}_N) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) + \Phi(-\widetilde{C}_N) \\
&= \Phi\left(c_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi(-c_N) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) + \Phi(-\widetilde{C}_N).
\end{aligned}$$

The third equality follows from the definitions of $\widetilde{C}_N$ and c_N. By mean value theorem,

$$\begin{aligned}
&\Phi\left(c_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi(-c_N) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) + \Phi(-\widetilde{C}_N) \\
&= \phi(t_1)(c_N - \widetilde{C}_N) + \phi(t_2)(c_N - \widetilde{C}_N), \text{for some numbers } t_1 \text{ and } t_2.
\end{aligned}$$

Also, if $\sqrt{N}\Delta/(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)$, $\widetilde{C}_N$ and c_N are bounded above, then $\phi(t_1) + \phi(t_2)$ is bounded away from zero. Note that $\phi(t_1) + \phi(t_2) > \phi(z_{\alpha/2}) = \alpha/2$. Hence, we have the following equivalence for any given $\eta > 0$

$$\mathbb{P}\left(c_N - \widetilde{C}_N > \frac{\eta}{\phi(t_1) + \phi(t_2)}\right) = \mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\Delta}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right).$$

Thus,

$$\mathbb{P}\left(c_N - \widetilde{C}_N > \eta\right) \leq \frac{2cN^{\frac{1}{2}-\upsilon}}{\eta\alpha\ \sqrt{2\pi}(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\ exp\left\{-\frac{N\Delta^2}{2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\} + \frac{1}{\sqrt{N}}.$$

Thus, the second difference

$$\mathbb{P}\left(\widehat{\psi}_l - c_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + c_N\widehat{\sigma}_u/\sqrt{N}\right) - \mathbb{P}\left(\widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N}\right)$$

is bounded above by

$$\mathbb{P}\left(c_N - \widetilde{C}_N > \eta\right) - 2\eta\ \theta,$$

where θ is the maximum value of the density of $\sqrt{N}(\widehat{\psi}_u - \psi_u)/\widehat{\sigma}_u$ and $\sqrt{N}(\widehat{\psi}_l - \psi_u)/\widehat{\sigma}_l$.

<u>Proof for the third difference</u>

The third difference can be re-written as

$$\begin{aligned}
&\mathbb{P}(\bar{C}_N < \widetilde{C}_N)\mathbb{P}\left(\widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \mid \bar{C}_N < \widetilde{C}_N\right)\\
&+ \mathbb{P}(\bar{C}_N < \widetilde{C}_N)\mathbb{P}\left(\widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N} \mid \bar{C}_N < \widetilde{C}_N\right)\\
&- \mathbb{P}(\bar{C}_N > \widetilde{C}_N)\mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \mid \bar{C}_N > \widetilde{C}_N\right)\\
&- \mathbb{P}(\bar{C}_N > \widetilde{C}_N)\mathbb{P}\left(\widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N} \mid \bar{C}_N > \widetilde{C}_N\right).
\end{aligned}$$

The probabilities are bounded above by 1. We just need to show that this difference is not too positive. Assuming that $\widehat{\psi}$ and $\widehat{\sigma}$ have uniformly continuous densities (to avoid high mass in a small area), it is sufficient to show the following

$$\mathbb{P}(\widetilde{C}_N - \bar{C}_N > \eta)$$

is bounded above. Define $\widehat{\Delta} = \widehat{\psi}_u - \widehat{\psi}_l$.

To prove the same, we will show that the following

$$\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right)$$

is bounded for a given $\eta > 0$.

Notice that we can break the event above into the following terms.

$$
\begin{aligned}
&\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta \\
&= \mathbb{1}\left(\frac{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}{\widehat{\sigma}_u \vee \widehat{\sigma}_l} \leq 1\right) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\} \\
&+ \mathbb{1}\left(\frac{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}{\widehat{\sigma}_u \vee \widehat{\sigma}_l} > 1\right) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\}.
\end{aligned}
$$

The goal is to show that the probability of this event is not too large. So we start by maximizing this probability and show that it is bounded. The first term has zero probability and hence, can be safely dropped. For the second term, notice the following

$$
\begin{aligned}
&\mathbb{1}\left\{\frac{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}{\widehat{\sigma}_u \vee \widehat{\sigma}_l} > 1, \, \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right\} \\
&\leq \mathbb{1}\left\{\phi\left(\frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) \times \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l} \left|\frac{\widehat{\sigma}_u \vee \widehat{\sigma}_l}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l} - 1\right| > \eta\right\},
\end{aligned}
$$

where the second bound follows by mean value theorem and the properties of the normal density ϕ. Now, by Markov's inequality we get the following bound for the expectation of the above term

$$
\mathbb{E}\left\{\frac{1}{\eta}\phi\left(\frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) \times \frac{\widehat{\Delta}\sqrt{N} \, |\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|}{(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)(\widehat{\sigma}_u \vee \widehat{\sigma}_l)}\right\} = \mathbb{E}\left[\frac{\sqrt{N} \, \widehat{\Delta} \, |\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|}{\eta\sqrt{2\pi}(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)} \, exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2}\right\}\right] \leq \frac{N^{\delta-\frac{1}{2}}}{\eta\sqrt{2\pi}(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)} \, e
$$

Also, notice that (see Das et al. (2021) for a detailed proof)

$$
\mathbb{E}\,|\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l| \leq \mathbb{E}\,|\widetilde{\sigma}_l - \widehat{\sigma}_l| \vee \mathbb{E}\,|\widetilde{\sigma}_u - \widehat{\sigma}_u| \lesssim \frac{1}{\sqrt{N}}.
$$

Thus,

$$
\begin{aligned}
&\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \eta\right) \\
&\leq \frac{\sqrt{N} \, \widehat{\Delta} \, |\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|}{\eta\sqrt{2\pi}(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)} \, exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2}\right\}.
\end{aligned}
$$

Next to show that $\mathbb{P}(\widetilde{C}_N - \bar{C}_N > \eta)$ is bounded above, we can use mean value theorem similar to what we did for the second difference.

Hence, we have that the following for any given $\eta > 0$

$$\mathbb{P}\left(\widetilde{C}_N - \bar{C}_N > \eta\right) = \mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \vee \widehat{\sigma}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right) > \frac{\eta\alpha}{2}\right).$$

Thus, the third difference

$$\mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N}\right) - \mathbb{P}\left(\widehat{\psi}_l - \widetilde{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \widetilde{C}_N\widehat{\sigma}_u/\sqrt{N}\right)$$

is bounded above by

$$\mathbb{P}\left(\widetilde{C}_N - \bar{C}_N > \eta\right) + 2\eta\ \theta,$$

where θ is the maximum value of the density of $\sqrt{N}(\widehat{\psi}_u - \psi_u)/\widehat{\sigma}_u$ and $\sqrt{N}(\widehat{\psi}_l - \psi_u)/\widehat{\sigma}_l$.

Combining the bounds

Now, we can show that the estimated confidence internal contains ψ_u with probability $1-\alpha$ and some error term that is not too negative. Similarly, one can show for ψ_l. And hence, the result follows for the target parameter ψ. For simplicity, we substitute Δ for $\psi_u - \psi_l$.

$$\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N}\right)\\
\leq\ & \left|\mathbb{P}\left(\frac{\widehat{\psi}_l - \psi_l}{\widehat{\sigma}_l/\sqrt{N}} \leq c_N + \frac{\Delta}{\widehat{\sigma}_l/\sqrt{N}}\right) - \Phi\left(c_N + \frac{\Delta}{\widehat{\sigma}_l/\sqrt{N}}\right)\right|\\
+\ & \left|\mathbb{P}\left(\frac{\widehat{\psi}_u - \psi_u}{\widehat{\sigma}_u/\sqrt{N}} \leq -c_N\right) - \Phi\left(-c_N\right)\right|\\
+\ & \Delta\mathbb{E}\left[\sqrt{N}\left|\frac{1}{\widehat{\sigma}_l} - \frac{1}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right| \frac{1}{\sqrt{2\pi}} exp\left\{-N\frac{\Delta^2}{2(\widehat{\sigma}_l \vee \widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}\right]\\
+\ & \mathbb{P}\left(c_N - \widetilde{C}_N > \eta\right) + 2\eta\ \theta + \mathbb{P}\left(\widetilde{C}_N - \bar{C}_N > \eta\right) + 2\eta\ \theta.
\end{aligned}$$

Now, if $\Delta = 0$, then in the context of this paper, this condition indicates that $\widehat{\Delta} = 0$. Thus, $\bar{C}_N = \widetilde{C}_N = c_N = z_{1-\alpha/2}$. Thus, the second and the third differences are zero when $\Delta = 0$. To incorporate this property here, we will use indicator terms.

$$
\begin{aligned}
&\leq \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{N}|\widehat{R}_{2,l}|}{\widetilde{\sigma}_l}\right) + \mathbb{E}\left\{\phi\left(\frac{\sqrt{N}\Delta}{\widehat{\sigma}_l \vee \widetilde{\sigma}_l}\right)\ \left(\frac{\Delta}{\widehat{\sigma}_l/\sqrt{N}} + c_N\right)\left|\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l} - 1\right|\right\} + C\mathbb{E}\left(\frac{\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}\right) \\
&+\ \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{N}|\widehat{R}_{2,u}|}{\widetilde{\sigma}_u}\right) + \frac{c_N}{\sqrt{2\pi}}\mathbb{E}\left(\left|\frac{\widehat{\sigma}_u}{\widetilde{\sigma}_u} - 1\right|\right) + C\mathbb{E}\left(\frac{\rho_u}{\widetilde{\sigma}_u^3\sqrt{N}}\right) \\
&+\ \Delta\mathbb{E}\left[\sqrt{N}\left|\frac{1}{\widehat{\sigma}_l} - \frac{1}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right| \frac{1}{\sqrt{2\pi}} exp\left\{-N\frac{\Delta^2}{2(\widehat{\sigma}_l \vee \widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}\right] \\
&+\ \mathbb{1}(\Delta \neq 0)\left[\frac{2cN^{\frac{1}{2}-\upsilon}}{\eta\alpha\sqrt{2\pi}(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\ exp\left\{-\frac{N\Delta^2}{2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}\right] + \mathbb{1}(\Delta \neq 0)\mathbb{E}\left(\frac{1}{\sqrt{N}}\right) \\
&+\ \mathbb{1}(\Delta \neq 0)\mathbb{E}\left[\frac{2\sqrt{N}\ \widehat{\Delta}\ |\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|}{\eta\alpha\ \sqrt{2\pi}(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)}\ exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2}\right\}\right] + \mathbb{1}(\Delta \neq 0)\mathbb{E}(4\eta\ \theta).
\end{aligned}
$$

We assume that $0 < \underline{\sigma} < \widehat{\sigma}_l, \widehat{\sigma}_u, \widetilde{\sigma}_l, \widetilde{\sigma}_u < \bar{\sigma} < \infty$.

If $\eta = N^{-\kappa}$ for some $\kappa > 0$, then the first two bounds (i.e., the first two differences computed at the beginning) decrease as N increases if the following holds

$$N > \frac{2(1/2 - \upsilon + \kappa)(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}{\Delta^2} \text{ and } \kappa + \delta < \frac{1}{2}.$$

To simplify the bound, let $\kappa = 1/2$. Also, by definition $\upsilon > 0$. Then the bound becomes

$$
\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \leq \psi_u \leq \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N}\right) \\
&\leq \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{N}|\widehat{R}_{2,l}|}{\widetilde{\sigma}_l}\right) + \mathbb{E}\left\{\phi\left(\frac{\sqrt{N}\Delta}{\widehat{\sigma}_l \vee \widetilde{\sigma}_l}\right)\ \left(\frac{\Delta}{\widehat{\sigma}_l/\sqrt{N}} + c_N\right)\left|\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l} - 1\right|\right\} + C\mathbb{E}\left(\frac{\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}}\right) \\
&+\ \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{N}|\widehat{R}_{2,u}|}{\widetilde{\sigma}_u}\right) + \frac{c_N}{\sqrt{2\pi}}\mathbb{E}\left(\left|\frac{\widehat{\sigma}_u}{\widetilde{\sigma}_u} - 1\right|\right) + C\mathbb{E}\left(\frac{\rho_u}{\widetilde{\sigma}_u^3\sqrt{N}}\right) \\
&+\ \Delta\mathbb{E}\left[\sqrt{N}\left|\frac{1}{\widehat{\sigma}_l} - \frac{1}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right| \frac{1}{\sqrt{2\pi}} exp\left\{-N\frac{\Delta^2}{2(\widehat{\sigma}_l \vee \widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}\right] \\
&+\ \mathbb{1}(\Delta \neq 0)\left[\frac{2cN^{1-\upsilon}}{\alpha\sqrt{2\pi}(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)}\ exp\left\{-\frac{N\Delta^2}{2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}\right\}\right] + \mathbb{1}(\Delta \neq 0)\mathbb{E}\left(\frac{1}{\sqrt{N}}\right) \\
&+\ \mathbb{1}(\Delta \neq 0)\mathbb{E}\left[\frac{\sqrt{2}N\ \widehat{\Delta}\ |\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|}{\alpha\sqrt{\pi}(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2(\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)}\ exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\sigma}_u \vee \widehat{\sigma}_l)^2}\right\}\right] + \mathbb{1}(\Delta \neq 0)\mathbb{E}\left(\frac{4\theta}{\sqrt{N}}\right).
\end{aligned}
$$

Further, using the inequality $e^{-w} < 1/w$ and $e^{-\frac{w}{2}} < \sqrt{3!/w^3}\ \forall\ w > 0$ all the exponential terms, and some rearrangement, we obtain the following simplified form.

$$
\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{\psi}_l - \bar{C}_N\widehat{\sigma}_l/\sqrt{N} \le \psi_u \le \widehat{\psi}_u + \bar{C}_N\widehat{\sigma}_u/\sqrt{N}\right) \\
\le\ & \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{N}|\widehat{R}_{2,l}|}{\widetilde{\sigma}_l} + \frac{\sqrt{N}|\widehat{R}_{2,u}|}{\widetilde{\sigma}_u}\right) + C\mathbb{E}\left(\frac{\rho_l}{\widetilde{\sigma}_l^3\sqrt{N}} + \frac{\rho_u}{\widetilde{\sigma}_u^3\sqrt{N}}\right) \\
&+\ \mathbb{E}\left\{\frac{1}{\sqrt{2\pi}}\ \left(\frac{\widehat{\sigma}_l \vee \widetilde{\sigma}_l}{\widehat{\sigma}_l} + c_N\right)\left|\frac{\widehat{\sigma}_l}{\widetilde{\sigma}_l} - 1\right|\right\} + \frac{c_N}{\sqrt{2\pi}}\mathbb{E}\left(\left|\frac{\widehat{\sigma}_u}{\widetilde{\sigma}_u} - 1\right|\right) \\
&+\ \mathbb{1}(\Delta \ne 0)\ \mathbb{E}\left\{\left|\frac{1}{\widehat{\sigma}_l} - \frac{1}{\widetilde{\sigma}_u \vee \widetilde{\sigma}_l}\right| \frac{\sqrt{2}(\widehat{\sigma}_l \vee \widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}{\sqrt{N\pi}\Delta}\right\} \\
&+\ \mathbb{1}(\Delta \ne 0)\ \mathbb{E}\left\{\frac{c\sqrt{2}\ \sqrt{6}\ (\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)^2}{\alpha\sqrt{\pi}N^{\frac{1}{2}+\upsilon}\ \Delta^3}\right\} + \mathbb{1}(\Delta \ne 0)\mathbb{E}\left\{\frac{1}{\sqrt{N}}\ (1+4\theta)\right\} \\
&+\ \mathbb{1}(\Delta \ne 0)\mathbb{E}\left\{\frac{2\sqrt{2}\ |\widetilde{\sigma}_u \vee \widetilde{\sigma}_l - \widehat{\sigma}_u \vee \widehat{\sigma}_l|}{\alpha\sqrt{\pi}\ \widehat{\Delta}\ (\widetilde{\sigma}_u \vee \widetilde{\sigma}_l)}\right\}.
\end{aligned}
$$

Further, let $0 < \underline{\sigma} \le \widehat{\sigma}_l, \widetilde{\sigma}_l, \widehat{\sigma}_u, \widetilde{\sigma}_u \le \bar{\sigma} < \infty$. Now using the bound $\mathbb{E}|\widehat{\sigma} - \widetilde{\sigma}| \lesssim N^{-\frac{1}{2}}$ for both the lower and the upper bound, we get the large sample bound as follows.

$$
\frac{1}{\sqrt{n}} + \sqrt{n}\mathbb{E}(|\widehat{R}_{2,l}| + \sqrt{N}|\widehat{R}_{2,u}|).
$$

$\square$

Proof of Proposition 2. One can rewrite the conditional risk ratio between lists 1 and 2 as follows:

$$\text{Risk ratio} = \frac{\mathbb{P}(Y_1 = 1 \mid Y_2 = 1, \mathbf{X} = \mathbf{x})}{\mathbb{P}(Y_1 = 1 \mid Y_2 = 0, \mathbf{X} = \mathbf{x})} = \frac{p_{12}(\mathbf{x})(1 - p_2(\mathbf{x}))}{(p_1(\mathbf{x}) - p_{12}(\mathbf{x}))p_2(\mathbf{x})} = \frac{q_{12}(\mathbf{x})\{1 - q_2(\mathbf{x})\gamma(\mathbf{x})\}}{(q_1(\mathbf{x}) - q_{12}(\mathbf{x}))q_2(\mathbf{x})\gamma(\mathbf{x})},$$

where we use the property that

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) = \frac{\mathbb{Q}(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x})}{\gamma(\mathbf{x})},$$

for $\mathbf{y} \neq \mathbf{0}$ and $\gamma(\mathbf{x}) = \mathbb{P}(\mathbf{Y} \neq \mathbf{0} | \mathbf{X} = \mathbf{x})$ is the conditional capture probability.

Let $\delta(\mathbf{x})$ denote the risk ratio function. Rearrangement of the expression of the risk ratio on the right gives us an expression for the conditional capture probability as follows.

$$\frac{1}{\gamma(\mathbf{x})} = \frac{\delta(\mathbf{x})\{q_1(\mathbf{x}) - q_{12}(\mathbf{x})\}q_2(\mathbf{x})}{q_{12}(\mathbf{x})} + q_2(\mathbf{x}).$$

□

Proof of Lemma 4.1.1. We derive the lower bound ψ_l^{-1} for ψ^{-1} in this proof. The steps for ψ_u^{-1} are similar.

We have seen before that

$$\psi_\delta^{-1} = \int \left[\frac{\{q_1(\mathbf{x}) - q_{12}(\mathbf{x})\}q_2(\mathbf{x})\delta(\mathbf{x})}{q_{12}(\mathbf{x})} + q_2(\mathbf{x}) \right] d\mathbb{Q}(\mathbf{x}) = \int \frac{1}{\gamma(\mathbf{x})} d\mathbb{Q}(\mathbf{x}).$$

Further, we need that $\gamma(\mathbf{X}) \leq 1$ for all $\mathbf{x}$.

Note that we can obtain ψ_l^{-1} by substituting the lowest possible value for $\delta(\mathbf{x})$ i.e., $1/\omega$ for each $\mathbf{x}$ so that $\gamma(\mathbf{x})$ is a valid probability. We ensure that it is a valid probability by using an indicator term for each $\mathbf{x}$. We define the following functionals

$$\begin{aligned}
\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} &= \left\{ \frac{q_1(\mathbf{x}) - q_{12}(\mathbf{x})}{\omega} + q_{12}(\mathbf{x}) \right\} \frac{q_2(\mathbf{x})}{q_{12}(\mathbf{x})} \\
\frac{1}{\gamma_{\omega}(\mathbf{x})} &= \left[\omega \{q_1(\mathbf{x}) - q_{12}(\mathbf{x})\} + q_{12}(\mathbf{x}) \right] \frac{q_2(\mathbf{x})}{q_{12}(\mathbf{x})}.
\end{aligned}$$

Thus, we obtain ψ_l^{-1} by integrating over $\gamma_{\frac{1}{\omega}}(\mathbf{x})^{-1}$ and ensuring that it is a valid probability using an indicator as follows. For ψ_u^{-1}, we substitute with $\gamma_\omega(\mathbf{x})^{-1}$.

$$\begin{aligned}
\psi_l^{-1} &= \int \left[\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1 \right\} + \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) > 1 \right\} \right] d\mathbb{Q}(\mathbf{x}) \\
&= \int \left[\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1 \right\} - \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1 \right\} + \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1 \right\} + \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) > 1 \right\} \right] d\mathbb{Q}(\mathbf{x}) \\
&= \int \left\{ \frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1 \right\} \mathbb{1}\left\{ \gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1 \right\} d\mathbb{Q}(\mathbf{x}) + 1.
\end{aligned}$$

□

Proof of theorem 4.2. We first derive the uncentered efficient influence function for ψ_l^{-1} assuming that we know the value the indicator $\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\}$ takes for each $\mathbf{x}$. Following that, we show that the derived function is the efficient influence function of ψ_l in general.

Note that, we can express ψ_l^{-1} as a function of the observed data distribution $\mathbb{Q}$, i.e., $\psi_l^{-1}(\mathbb{Q})$. We use the notation $\psi_l^{-1}(\mathbb{Q}, \bar{\mathbb{Q}})$ to denote the following functional

$$\psi_l^{-1}(\mathbb{Q}, \bar{\mathbb{Q}}) = \int \left\{\frac{1}{\omega}\frac{q_1(\mathbf{x})q_2(\mathbf{x})}{q_{12}(\mathbf{x})} + \left(1 - \frac{1}{\omega}\right) q_2(\mathbf{x}) - 1\right\} \mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} q(\mathbf{x}) d\mathbf{x} + 1,$$
$$\text{where } \frac{1}{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x})} = \frac{1}{\omega}\frac{\bar{q}_1(\mathbf{x})\bar{q}_2(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})} + \left(1 - \frac{1}{\omega}\right) \bar{q}_2(\mathbf{x}) + 1.$$

We assume that, we know $\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\}$ and derive the efficient influence function for $\psi_l^{-1}(\mathbb{Q}, \mathbb{Q})$.

To find a candidate influence function, we consider a special parametric submodel (i.e., deviation from $\mathbb{Q}$) given by $\mathbb{Q}_\epsilon$ with density $q_\epsilon = (1-\epsilon)q(\mathbf{z}) + \epsilon\bar{q}(\mathbf{z})$ where $\bar{q} = \bar{q}(\mathbf{z}) = \mathbb{1}(\mathbf{z} = \tilde{\mathbf{z}})$ is a point mass at $\mathbf{Z} = \tilde{\mathbf{z}}$, and for which the pathwise derivative

$$\frac{\partial}{\partial \epsilon}\left\{\frac{1}{\psi_l(\mathbb{Q}_\epsilon, \mathbb{Q})}\right\}\Bigg|_{\epsilon=0}$$

actually equals the influence function (in the discrete case) Mises (1947); Hampel (1974). We also let $q_{s,\epsilon}(\mathbf{x})$ denote the analog of $q_s(\mathbf{x})$ under the submodel for $s \in \{1, 2, 12\}$, e.g., the marginal density for $\mathbf{X}$ under $\mathbb{Q}_\epsilon$ is

$$q_\epsilon(\mathbf{x}) = \sum_{\mathbf{y}} q_\epsilon(\mathbf{z}) = (1-\epsilon)q(\mathbf{x}) + \epsilon\mathbb{1}(\mathbf{x} = \tilde{\mathbf{x}}).$$

Now the above pathwise derivative equals

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\left\{\frac{1}{\psi(\mathbb{Q}_\epsilon, \mathbb{Q})}\right\}\Bigg|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\int \left\{\frac{1}{\omega}\frac{q_{2,\epsilon}(\mathbf{x})q_{1,\epsilon}(\mathbf{x})}{q_{12,\epsilon}(\mathbf{x})} + \left(1 - \frac{1}{\omega}\right) q_{2,\epsilon}(\mathbf{x}) - 1\right\} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} q_\epsilon(\mathbf{x}) d\mathbf{x}\Bigg|_{\epsilon=0} \\
&= \int \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{\partial}{\partial \epsilon}\left\{\frac{q_{2,\epsilon}(\mathbf{x})q_{1,\epsilon}(\mathbf{x})}{\omega\ q_{12,\epsilon}(\mathbf{x})} q_\epsilon(\mathbf{x}) + \left(1 - \frac{1}{\omega}\right) q_{2,\epsilon}(\mathbf{x}) q_\epsilon(\mathbf{x})\right\} d\mathbf{x}\Bigg|_{\epsilon=0} \\
&= \int \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{q_{2,\epsilon}(\mathbf{x})q_{1,\epsilon}(\mathbf{x})}{\omega\ q_{12,\epsilon}(\mathbf{x})} q_\epsilon(\mathbf{x}) \left\{\frac{q'_{1,\epsilon}(\mathbf{x})}{q_{1,\epsilon}(\mathbf{x})} + \frac{q'_{2,\epsilon}(\mathbf{x})}{q_{2,\epsilon}(\mathbf{x})} - \frac{q'_{12,\epsilon}(\mathbf{x})}{q_{12,\epsilon}(\mathbf{x})} + \frac{q'_\epsilon(\mathbf{x})}{q_\epsilon(\mathbf{x})}\right\} d\mathbf{x}\Bigg|_{\epsilon=0} \\
&\quad + \int \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left(1 - \frac{1}{\omega}\right) q_{2,\epsilon}(\mathbf{x})\ q_\epsilon(\mathbf{x}) \left\{\frac{q'_{2,\epsilon}(\mathbf{x})}{q_{2,\epsilon}(\mathbf{x})} + \frac{q'_\epsilon(\mathbf{x})}{q_\epsilon(\mathbf{x})}\right\} d\mathbf{x}\Bigg|_{\epsilon=0}.
\end{aligned}$$

where the last equality follows by the product rule. For the discrete case, we use the notation of the integral to denote summation over $\mathbf{x}$.

For the derivatives appearing above, by the definition of $\bar{q}$, we have $q'_\epsilon(\mathbf{x}) = \frac{\partial}{\partial \epsilon} q_\epsilon(\mathbf{x}) = \mathbb{1}(\mathbf{x} = \tilde{\mathbf{x}}) - q(\mathbf{x})$. Similarly, by using derivative of product rule for $q'_{1,\epsilon}(\mathbf{x})$, we get

$$\begin{aligned}
q'_{1,\epsilon}(\mathbf{x}) = \frac{\partial}{\partial \epsilon} q_{1,\epsilon}(\mathbf{x}) &= \frac{\partial}{\partial \epsilon} \frac{\mathbb{Q}_\epsilon(Y_1 = 1, \mathbf{X} = \mathbf{x})}{q_\epsilon(\mathbf{x})} \\
&= \frac{\partial}{\partial \epsilon} \frac{(1-\epsilon)\mathbb{Q}(Y_1 = 1, \mathbf{X} = \mathbf{x}) + \epsilon \mathbb{1}(\tilde{Y}_1 = 1, \mathbf{x} = \tilde{\mathbf{x}})}{(1-\epsilon) q(\mathbf{x}) + \epsilon \mathbb{1}(\mathbf{x} = \tilde{\mathbf{x}})}, \text{ where } \tilde{Y}_1 = \mathbb{1}(\tilde{y}_1 = 1) \\
&= \frac{-\mathbb{Q}(Y_1 = 1, \mathbf{X} = \mathbf{x}) + \mathbb{1}(\tilde{Y}_1 = 1, \mathbf{x} = \tilde{\mathbf{x}})}{(1-\epsilon) q(\mathbf{x}) + \epsilon \mathbb{1}(\mathbf{x} = \tilde{\mathbf{x}})} \\
&\quad - \frac{(1-\epsilon)\mathbb{Q}(Y_1 = 1, \mathbf{X} = \mathbf{x}) + \epsilon \mathbb{1}(\tilde{Y}_1 = 1, \mathbf{x} = \tilde{\mathbf{x}})}{\{(1-\epsilon) q(\mathbf{x}) + \epsilon \mathbb{1}(\mathbf{x} = \tilde{\mathbf{x}})\}^2} q'_\epsilon(\mathbf{x}).
\end{aligned}$$

The last step follows from the product rule of derivatives. Finally, setting $\epsilon = 0$, we get $q'_{1\epsilon}(\mathbf{x})|_{\epsilon=0} = \frac{\mathbb{1}(\mathbf{x}=\tilde{\mathbf{x}})}{q(\mathbf{x})}\{\tilde{Y}_1 - q_1(\mathbf{x})\}$. The derivatives for $q_{2,\epsilon}$ and $q_{12,\epsilon}$ follow similarly.

Thus, combining the above results and using the discrete nature of the distribution, we get

$$\begin{aligned}
\phi_l(\tilde{\mathbf{x}}, \tilde{\mathbf{Y}}; \mathbb{Q}) &= \frac{\partial}{\partial \epsilon} \left\{ \frac{1}{\psi(\mathbb{Q}_\epsilon, \mathbb{Q})} \right\} \Bigg|_{\epsilon = 0} \\
&= \sum_{\mathbf{x}} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{q_2(\mathbf{x}) q_1(\mathbf{x})}{\omega\ q_{12}(\mathbf{x})} \mathbb{1}(\mathbf{x} = \widetilde{\mathbf{x}}) \left\{ \frac{\widetilde{Y}_1 - q_1(\mathbf{x})}{q_1(\mathbf{x})} + \frac{\widetilde{Y}_2 - q_2(\mathbf{x})}{q_2(\mathbf{x})} - \frac{\widetilde{Y}_1 \widetilde{Y}_2 - q_{12}(\mathbf{x})}{q_{12}(\mathbf{x})} \right\} \\
&\quad + \sum_{\mathbf{x}} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{q_2(\mathbf{x}) q_1(\mathbf{x})}{\omega\ q_{12}(\mathbf{x})} \{\mathbb{1}(\mathbf{x} = \widetilde{\mathbf{x}}) - q(\mathbf{x})\} \\
&\quad + \sum_{\mathbf{x}} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left(1 - \frac{1}{\omega}\right) q_2(\mathbf{x}) \left\{ \mathbb{1}(\mathbf{x} = \widetilde{\mathbf{x}})\ \frac{\widetilde{Y}_2 - q_2(\mathbf{x})}{q_2(\mathbf{x})} + \mathbb{1}(\mathbf{x} = \widetilde{\mathbf{x}}) - q(\mathbf{x}) \right\} \\
&= \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\widetilde{\mathbf{x}}) \leq 1\right\} \frac{q_2(\widetilde{\mathbf{x}}) q_1(\widetilde{\mathbf{x}})}{\omega\ q_{12}(\widetilde{\mathbf{x}})} \left\{ \frac{\widetilde{Y}_1}{q_1(\widetilde{\mathbf{x}})} + \frac{\widetilde{Y}_2}{q_2(\widetilde{\mathbf{x}})} - \frac{\widetilde{Y}_1 \widetilde{Y}_2}{q_{12}(\widetilde{\mathbf{x}})} \right\} \\
&\quad - \sum_{\mathbf{x}} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{q_2(\mathbf{x}) q_1(\mathbf{x})}{\omega\ q_{12}(\mathbf{x})} q(\mathbf{x}) - \sum_{\mathbf{x}} \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left(1 - \frac{1}{\omega}\right) q_2(\mathbf{x}) q(\mathbf{x}) \\
&\quad + \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\widetilde{\mathbf{x}}) \leq 1\right\} \left(1 - \frac{1}{\omega}\right) \widetilde{Y}_2 \\
&= \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\widetilde{\mathbf{x}}) \leq 1\right\} \left[\frac{q_2(\widetilde{\mathbf{x}}) q_1(\widetilde{\mathbf{x}})}{\omega\ q_{12}(\widetilde{\mathbf{x}})} \left\{ \frac{\widetilde{Y}_1}{q_1(\widetilde{\mathbf{x}})} + \frac{\widetilde{Y}_2}{q_2(\widetilde{\mathbf{x}})} - \frac{\widetilde{Y}_1 \widetilde{Y}_2}{q_{12}(\widetilde{\mathbf{x}})} \right\} + \left(1 - \frac{1}{\omega}\right) \widetilde{Y}_2 - \frac{1}{\gamma_{\frac{1}{\omega}}(\widetilde{\mathbf{x}})} \right].
\end{aligned}$$

The third equality follows from the definition of $\gamma_{\frac{1}{\omega}}$ and rearrangement of the terms.

Next, we will show that under the margin condition in assumption 5, the above function ϕ_l is also the efficient influence function of $\psi_l(\mathbb{Q}_\epsilon, \mathbb{Q}_\epsilon)^{-1}$, i.e., when the indicator is also not known and is subject to fluctuations. To show the same, we will evaluate the remainder term below and show that it is of second order under assumption 5. First, note that ϕ_l can be expressed as follows

$$\phi_l(\tilde{\mathbf{x}}, \tilde{\mathbf{Y}}; \mathbb{Q}) = \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\widetilde{\mathbf{x}}) \leq 1\right\} \left[\frac{q_2(\widetilde{\mathbf{x}}) q_1(\widetilde{\mathbf{x}})}{\omega\ q_{12}(\widetilde{\mathbf{x}})} \left\{\frac{\widetilde{Y}_1}{q_1(\widetilde{\mathbf{x}})} + \frac{\widetilde{Y}_2}{q_2(\widetilde{\mathbf{x}})} - \frac{\widetilde{Y}_1\widetilde{Y}_2}{q_{12}(\widetilde{\mathbf{x}})}\right\} + \left(1-\frac{1}{\omega}\right)\widetilde{Y}_2 - 1\right] + 1 - \psi_l^{-1}.$$

The remainder term for distribution $\mathbb{Q}$ and some other distribution $\bar{\mathbb{Q}}$ is as follows:

$$\begin{aligned}
&R_{2,l}(\mathbb{Q}, \bar{\mathbb{Q}}) = \int \bar{\phi}_l(\mathbf{z}, \bar{\mathbb{Q}}) d\mathbb{Q}(\mathbf{z}) + \bar{\psi}_l^{-1} - \psi_l^{-1} \\
&= \int \left(\mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left[\frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\omega\ \bar{q}_{12}(\mathbf{x})} \left\{\frac{q_1(\mathbf{x})}{\bar{q}_1(\mathbf{x})} + \frac{q_2(\mathbf{x})}{\bar{q}_2(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})}\right\} + \left(1 - \frac{1}{\omega}\right) q_2(\mathbf{x}) - 1\right] + 1 - \psi_l^{-1} \right) d\mathbb{Q}(\mathbf{x}) \\
&= \int \left(\mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left[\frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\omega\ \bar{q}_{12}(\mathbf{x})} \left\{\frac{q_1(\mathbf{x})}{\bar{q}_1(\mathbf{x})} + \frac{q_2(\mathbf{x})}{\bar{q}_2(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})}\right\} + \left(1 - \frac{1}{\omega}\right) q_2(\mathbf{x}) - 1\right] \right. \\
&\quad \left. - \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1\right\} \right) d\mathbb{Q}(\mathbf{x}) \\
&= \int \left(\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left[\frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\omega\ \bar{q}_{12}(\mathbf{x})} \left\{\frac{q_1(\mathbf{x})}{\bar{q}_1(\mathbf{x})} + \frac{q_2(\mathbf{x})}{\bar{q}_2(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})}\right\} + \left(1 - \frac{1}{\omega}\right) q_2(\mathbf{x}) - \frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})}\right] \right. \\
&\quad \left. + \left[\mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} - \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\}\right] \left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1\right\} \right) d\mathbb{Q}(\mathbf{x}) \\
&= \int \mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \left[\frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\omega\ \bar{q}_{12}(\mathbf{x})} \left\{\frac{q_1(\mathbf{x})}{\bar{q}_1(\mathbf{x})} + \frac{q_2(\mathbf{x})}{\bar{q}_2(\mathbf{x})} - \frac{q_{12}(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})}\right\} + \frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\omega\ \bar{q}_{12}(\mathbf{x})}\right] d\mathbb{Q}(\mathbf{x}) \\
&\quad + \int \left[\mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} - \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\}\right] \left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1\right\} d\mathbb{Q}(\mathbf{x}) \\
&= \int \mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \frac{1}{\omega \bar{q}_{12}(\mathbf{x})} \Bigg[\{q_1(\mathbf{x}) - \bar{q}_1(\mathbf{x})\} \{\bar{q}_2(\mathbf{x}) - q_2(\mathbf{x})\} \\
&\quad + \{q_{12}(\mathbf{x}) - \bar{q}_{12}(\mathbf{x})\} \left\{\frac{q_2(\mathbf{x})q_1(\mathbf{x})}{q_{12}(\mathbf{x})} - \frac{\bar{q}_2(\mathbf{x})\bar{q}_1(\mathbf{x})}{\bar{q}_{12}(\mathbf{x})}\right\} \Bigg] d\mathbb{Q}(\mathbf{x}) \\
&\quad + \int \left[\mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} - \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\}\right] \left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1\right\} d\mathbb{Q}(\mathbf{x}).
\end{aligned}$$

The first term is in a multiplicative form and easy to bound above. For the second term, we follow the approaches from Bonvini and Kennedy (2020) and Kennedy et al. (2020). Below is the bound.

$$\begin{aligned}
&\int \left| \mathbb{1}\left\{\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} - \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{x}) \leq 1\right\} \right| \left| \frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1 \right| d\mathbb{Q}(\mathbf{x}) \\
&\leq \mathbb{1}\left\{ \left|\gamma_{\frac{1}{\omega}}(\mathbf{x}) - 1\right| \leq \left|\bar{\gamma}_{\frac{1}{\omega}}(\mathbf{x}) - \gamma_{\frac{1}{\omega}}(\mathbf{x})\right| \right\} \left| \frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{x})} - 1 \right| d\mathbb{Q}(\mathbf{x}) \\
&\lesssim \frac{1}{\epsilon} \left\| \bar{\gamma}_{\frac{1}{\omega}} - \gamma_{\frac{1}{\omega}} \right\|_{\infty}^{1+\nu},
\end{aligned}$$

since $q_{12} \wedge \bar{q}_{12} \geq \epsilon$ implies $\gamma_{\frac{1}{\omega}}(\mathbf{x}) \geq q_{12}(\mathbf{x}) \geq \epsilon$. This shows that $\mathbb{E}|\widehat{R}_{2,l}|$ is of second-order when $\nu = 1$ in the margin condition. Hence, ψ_l is the efficient influence function of ψ_l^{-1}.

Similarly, the corresponding efficient influence function for ψ_u^{-1} is

$$\begin{aligned}\phi_l(\mathbf{x}, \mathbf{Y}; \mathbb{Q}) = \mathbb{1}\{\gamma_\omega(\mathbf{x}) \leq 1\} &\left[\frac{\omega\ q_2(\mathbf{x}) q_1(\mathbf{x})}{q_{12}(\mathbf{x})} \left\{\frac{Y_1}{q_1(\mathbf{x})} + \frac{Y_2}{q_2(\mathbf{x})} - \frac{Y_1 Y_2}{q_{12}(\mathbf{x})}\right\} + (1-\omega)\, Y_2 - 1\right] \\ &+ 1 - \psi_l^{-1}.\end{aligned}$$

□

Proof of corollary 4.2.1. We derive the variance of ϕ_l, which is also the efficiency bound. First, note that the efficient influence function can be represented as follows

$$\begin{aligned}\phi_l(\mathbf{Z}; \mathbb{Q}) = \mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\widetilde{\mathbf{x}}) \leq 1\right\} &\left(\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})}\left[\frac{(Y_1 - Y_1Y_2)/\omega + Y_1Y_2}{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}/\omega + q_{12}(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\right] - 1\right) \\ &+ 1 - \psi_l^{-1}.\end{aligned}$$

Then we use the formula for total variance by conditioning on the covariates $\mathbf{X}$ as follows.

$$var\,(\phi_l) = \mathbb{E}\left\{var\,(\phi_l|\mathbf{X})\right\} + var\left\{\mathbb{E}\,(\phi_l|\mathbf{X})\right\}.$$

We use following the variance and covariance formulas for the indicator terms Y_1 and Y_2.

$$\begin{aligned}var(Y_1 - Y_1Y_2 \mid \mathbf{X}) &= \{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}\{1 - q_1(\mathbf{X}) + q_{12}(\mathbf{X})\}. \\ var\{(Y_1 - Y_1Y_2)/\omega + Y_1Y_2 \mid \mathbf{X}\} &= \{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}\{1 - q_1(\mathbf{X}) + q_{12}(\mathbf{X})\}/\omega^2 \\ &\quad + q_{12}(\mathbf{X})\{1 - q_{12}(\mathbf{X})\} - 2\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}q_{12}(\mathbf{X})/\omega \\ &= \{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}/\omega^2 + q_{12}(\mathbf{X}) - [\{q_1(\mathbf{X}) - q_2(\mathbf{X})\}/\omega + q_{12}(\mathbf{X})]^2. \\ \mathrm{cov}(Y_1 - Y_1Y_2, Y_1Y_2 \mid \mathbf{X}) &= -\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}q_{12}(\mathbf{X}). \\ \mathrm{cov}(Y_1 - Y_1Y_2, Y_2 \mid \mathbf{X}) &= -\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}q_2(\mathbf{X}).\end{aligned}$$

Using conditional variance, we get

$$\begin{aligned}&\mathbb{E}\left(\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\} \times \frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})^2} var\left[\frac{(Y_1 - Y_1Y_2)/\omega + Y_1Y_2}{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}/\omega + q_{12}(\mathbf{X})} + \frac{Y_2}{q_2(\mathbf{X})} - \frac{Y_1Y_2}{q_{12}(\mathbf{X})}\middle|\mathbf{X}\right]\right) \\ &+ var\left[\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\}\right].\end{aligned}$$

After further simplification we get the following.

$$\mathbb{E}\left(\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\left[\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})}\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\}\left\{\frac{1}{\omega q_{12}(\mathbf{X})} - 1\right\} + \frac{q_0(\mathbf{X})}{\omega \gamma_{\frac{1}{\omega}}(\mathbf{X}) q_{12}(\mathbf{X})}\right.\right.$$
$$\left.\left. + \left(1 - \frac{1}{\omega}\right)\frac{\{q_1(\mathbf{X}) - q_{12}(\mathbf{X})\}^2 q_2(\mathbf{X})^2}{\omega^2 q_{12}(\mathbf{X})^3}\right]\right) + var\left[\left\{\frac{1}{\gamma_{\frac{1}{\omega}}(\mathbf{X})} - 1\right\}\mathbb{1}\left\{\gamma_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right].$$

□

Proof of theorem 2.2. The error in estimation for the proposed estimator is approximated by a sample average as follows.

$$\begin{aligned}\mathcal{E}_N = \widehat{\psi}^{-1}_{l,proposed} - \psi_l^{-1} - \mathbb{Q}_N \phi_l &= \mathbb{Q}_N \widehat{\phi}_l + \widehat{\psi}^{-1}_{l,pi} - \psi_l^{-1} - \mathbb{Q}_N \phi_l \\ &= \mathbb{Q}_N(\widehat{\phi}_l - \phi_l) - \mathbb{Q}\widehat{\phi}_l + \widehat{R}_{2,l} \\ &= (\mathbb{Q}_N - \mathbb{Q})(\widehat{\phi}_l - \phi_l) + \widehat{R}_{2,l}, \text{ since } \mathbb{Q}\phi_l = 0.\end{aligned}$$

The second equality follows using the property of efficient influence function as follows

$$\widehat{\psi}_l^{-1} - \psi_l^{-1} = -\int \widehat{\phi}_l(\mathbf{z}) d\mathbb{Q}(\mathbf{z}) + \widehat{R}_{2,l} = -\mathbb{Q}\widehat{\phi}_l + \widehat{R}_{2,l}.$$

The first term is converging to 0 since it is a sample average minus the population average of i.i.d. terms.

Nest, we will show that $\mathbb{E}(\mathcal{E}_N^2)$ is bounded. By the law of conditional expectation,

$$\mathbb{E}(\mathcal{E}_N) = \mathbb{E}\{\mathbb{E}(\mathcal{E}_N | \mathbf{Z}^n)\} = \mathbb{E}\left[\mathbb{E}\left\{(\mathbb{Q}_N - \mathbb{Q})\left(\widehat{\phi}_l - \phi_l\right) + \widehat{R}_{2,l} \middle| \mathbf{Z}^n\right\}\right].$$

The first term has expectation because it is the difference of sample average and population average. Further, $\widehat{R}_{2,l}$ is a function is the training sample, and thus

$$\mathbb{E}(\mathcal{E}_N) = \mathbb{E}(\widehat{R}_{2,l}).$$

Next, the variance of $\mathcal{E}_N$ is

$$\begin{aligned} var(\mathcal{E}_N) &= var\{\mathbb{E}(\mathcal{E}_N | \mathbf{Z}^n)\} + \mathbb{E}\{var(\mathcal{E}_N | \mathbf{Z}^n)\} \\ &= var\left(\widehat{R}_{2,l}\right) + \mathbb{E}\left[var\left\{(\mathbb{Q}_N - \mathbb{Q})\left(\widehat{\phi}_l - \phi_l\right) \middle| \mathbf{Z}^n\right\}\right] \\ &\leq var\left(\widehat{R}_{2,l}\right) + \mathbb{E}\left(\frac{1}{N}\left\|\widehat{\phi}_l - \phi_l\right\|^2\right),\end{aligned}$$

where the second step follows via simple triangle inequality. For a more detailed discussion, we refer to Das et al. (2021). Finally, we have that

$$\mathbb{E}(\mathcal{E}_N^2) = \mathbb{E}(\widehat{R}_{2,l}^2) + \mathbb{E}\left(\frac{1}{N}\left\|\widehat{\phi}_l - \phi_l\right\|^2\right).$$

□

Proof of Theorem 4.5. We present the proof for the estimated lower bound of n i.e., $\widehat{n}_l$. The proof for $\widehat{n}_u$ follows similarly. We have defined $\widehat{n}_l = N\widehat{\psi}_l^{-1}$, which depends on two random quantities (i) the number of observations N and (ii) the estimator of the lower bound of the inverse capture probability $\widehat{\psi}_l^{-1}$.

Calculation of mean and variance of $\widehat{n}$

First, we re-write $N\widehat{\psi}_l^{-1}$ as a sample average for ease of calculation. As mentioned in the proof of Theorem 2.2, we use $\mathbb{Q}\left[\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right] + 1$ to denote the population average conditioned on training sample $\mathbf{Z}^n$ i.e., $\int \widehat{\varphi}(\mathbf{z})\mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\} d\mathbb{Q}(\mathbf{z}) + 1$.

$$\begin{aligned}
\widehat{n}_l =&\ N\widehat{\psi}_l^{-1} - N\psi_l^{-1} + N\psi_l^{-1} \\
=&\ N\left((\mathbb{Q}_N - \mathbb{Q})\left[\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right] + \widehat{R}_{2,l}\right) + N\psi_l^{-1} \\
=&\ n\mathbb{P}_n \underbrace{\left[\mathbb{1}(\mathbf{Y} \neq \mathbf{0})\left(\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\} - \mathbb{Q}\left[\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\right]\right) + \mathbb{1}(\mathbf{Y} \neq \mathbf{0})\left(\widehat{R}_{2,l} + \psi_l^{-1}\right)\right]}_{\zeta},
\end{aligned}$$

where the last equality follows using $N = n\mathbb{P}_n\mathbb{1}(\mathbf{Y} \neq \mathbf{0})$. Thus, $\widehat{n}_l$ is a sample average $n\mathbb{P}_n\zeta$. Moreover, when we condition on the training sample $\mathbf{Z}^n$, the ζ's are i.i.d. We present the conditional mean and variance of ζ below.

$$\mathbb{E}(\zeta|\mathbf{Z}^n) = \mathbb{E}\{\mathbb{E}(\zeta|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n)|\mathbf{Z}^n\} = \psi(\widehat{R}_{2,l} + \psi_l^{-1}).$$

$$\begin{aligned}
var(\zeta|\mathbf{Z}^n) &= var\{\mathbb{E}(\zeta|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n)\} + \mathbb{E}\{var(\zeta|\mathbf{Y} \neq \mathbf{0}, \mathbf{Z}^n)\} \\
&= var\left[\mathbb{1}(\mathbf{Y} \neq \mathbf{0})(\widehat{R}_{2,l} + \psi_l^{-1})\middle|\mathbf{Z}^n\right] + \mathbb{E}\left(\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) var\left[\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\middle|\mathbf{Z}^n\right]\middle|\mathbf{Z}^n\right) \\
&= \psi(1-\psi)(\widehat{R}_{2,l} + \psi_l^{-1})^2 + \psi\ var\left[\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\middle|\mathbf{Z}^n\right] \\
&= \frac{\psi(1-\psi)}{\psi_l^2}(\psi_l\widehat{R}_{2,l} + 1)^2 + \psi\widetilde{\varsigma}_l^2, \quad \text{defining } \widetilde{\varsigma}_l^2 = var\left[\widehat{\varphi}_l \mathbb{1}\left\{\widehat{\gamma}_{\frac{1}{\omega}}(\mathbf{X}) \leq 1\right\}\middle|\mathbf{Z}^n\right].
\end{aligned}$$

The population expectation of $\widehat{n}$ is $n\psi\mathbb{E}(\widehat{R}_{2,l}) + n\psi\psi_l^{-1}$. And the population variance is presented below.

$$\begin{aligned}
var(\widehat{n}_l) &= var\{\mathbb{E}(\widehat{n}_l|\mathbf{Z}^n)\} + \mathbb{E}\{var(\widehat{n}_l|\mathbf{Z}^n)\} \\
&= var\{\mathbb{E}(n\mathbb{P}_n\zeta|\mathbf{Z}^n)\} + \mathbb{E}\{var(n\mathbb{P}_n\zeta|\mathbf{Z}^n)\} \\
&= var(n\psi\widehat{R}_{2,l} + n\psi\psi_l^{-1}) + \mathbb{E}\{n\ var(\zeta|\mathbf{Z}^n)\} \\
&= n^2\psi^2 var(\widehat{R}_{2,l}) + n\psi\mathbb{E}\left(\widehat{\varsigma}_l^2\right) + n\frac{\psi(1-\psi)}{\psi_l^2}\mathbb{E}(\psi_l\widehat{R}_{2,l}+1)^2.
\end{aligned}$$

Thus, using the definition $n_l \equiv n\psi\psi_l^{-1}$, we get $\mathbb{E}(\widehat{n}_l - n_l)^2 = n^2\psi^2\mathbb{E}(\widehat{R}_{2,l}^2) + n\psi\mathbb{E}\left(\widehat{\varsigma}_l^2\right) + n\frac{\psi(1-\psi)}{\psi_l^2}\mathbb{E}(\psi_l\widehat{R}_{2,l}+1)^2$. If $\mathbb{E}|\widehat{R}_{2,l}|$ is sufficiently small, then $\widehat{n}_l - n_l$ has expectation approximately 0 and variance approximately $n\psi\varsigma_l^2 + n\psi(1-\psi)/\psi_l^2$, where $\varsigma_l^2 = var\left[\varphi_l\mathbb{1}\left\{\zeta(\mathbf{X}) \le \frac{1}{\omega}\right\}\right]$. □

Proof of Theorem 4.5. We present the proof for the estimated lower limit of ψ i.e., $\widehat{n}_u$. The proof for $\widehat{n}_l$ follows similarly. We ultimately want to see the error in the coverage of the following Imbens and Manski (2004) $(1-\alpha) \times 100\%$ confidence interval

$$\left[\widehat{n}_l - \bar{C}_N \sqrt{N}\widehat{\tau}_l,\ \widehat{n}_u + \bar{C}_N \sqrt{N}\widehat{\tau}_u\right].$$

Below we evaluate the finite sample coverage error of this estimated confidence interval. The error in coverage is

$$\left|(1-\alpha) - \mathbb{P}\left(\widehat{n}_l - \bar{C}_N \sqrt{N}\widehat{\tau}_l \leq n \leq \widehat{n}_u + \bar{C}_N \sqrt{N}\widehat{\tau}_u\right)\right|.$$

Note that $\bar{C}_N$ satisfies

$$\Phi\left(\bar{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi(-\bar{C}_N) = 1-\alpha.$$

The quantity $\bar{C}_N$ is a stochastic quantity that depends on $\widehat{n}_l$, $\widehat{n}_u$, $\sqrt{N}\widehat{\tau}_l$ and $\sqrt{N}\widehat{\tau}_u$. Hence, to apply Berry-Esseen bound, we use a non-stochastic approximation of $\bar{C}_N$, which is c_N. c_N is a constant for a given n (also N) and training data $\mathbf{Z}^n$ unlike $\bar{C}_N$, and satisfies

$$\Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) - \Phi(-c_N) = 1-\alpha.$$

We also define an intermediate quantity $\widetilde{C}_N$ which satisfies

$$\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi(-\widetilde{C}_N) = 1-\alpha.$$

Remark 29. *In the context of this paper, the difference* $n_u - n_l$ *is zero implies that the* $\widetilde{\tau}_u = \widetilde{\tau}_l$ *and also* $\widehat{n}_u = \widehat{n}_l$. *Thus, when the difference is zero,* $\bar{C}_N = \widetilde{C}_N = c_N = z_{1-\alpha/2}$.

Next, following the approach of Imbens and Manski (2004), we show that the estimated confidence interval contains n_u with probability $1-\alpha$ and some additional error. The proof for n_l follows similarly. Since, ψ lies between n_u and n_l, the estimated interval will contain ψ with probability $1-\alpha$ and some additional error which is the maximum of the errors of n_u and n_l.

Below we evaluate the error in coverage of n_u by the estimated confidence interval.

$$
\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u\right) \\
&= (1-\alpha) - \mathbb{P}\left(\widehat{n}_l - c_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + c_N\sqrt{N}\widehat{\tau}_u\right) && \text{(B.1)} \\
&+ \mathbb{P}\left(\widehat{n}_l - c_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + c_N\sqrt{N}\widehat{\tau}_u\right) - \mathbb{P}\left(\widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u\right) && \text{(B.2)} \\
&+ \mathbb{P}\left(\widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u\right) - \mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u\right). && \text{(B.3)}
\end{aligned}
$$

We need to show that the three differences are either positive or close to zero with probability close to 1.

Before showing the above, we first derive the bounds on the differences between the estimated and the population variances. We want to show that $\sqrt{N}\widehat{\tau}_l - \sqrt{n\psi}\widetilde{\tau}_l$ and $\sqrt{N}\widehat{\tau}_u - \sqrt{n\psi}\widetilde{\tau}_u$ are not too negative.

Bound on the differences among the variance terms

In this part, we will derive upper bounds on the following quantities

1. $\mathbb{E}\left(1 - \frac{\widehat{\tau}_l\sqrt{N}}{\widetilde{\tau}_l\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)$
2. $\mathbb{E}\left(1 - \frac{\widehat{\tau}_u\sqrt{N}}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)$
3. $\mathbb{E}\left(1 - \frac{\widehat{\tau}_l\sqrt{N}}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)$.

These bounds are required when we evaluate the bounds on differences B.1 and B.3 in the later parts of the proof. Below are the derivations.

1. We expand the following difference using the previously stated definitions:

$$
\begin{aligned}
\widehat{\tau}_l^2 N - \widetilde{\tau}_l^2 n\psi &= \widehat{\tau}_l^2 N - \widehat{\tau}_l^2 n\psi + \widehat{\varsigma}_l^2 n\psi + \frac{1-\widehat{\psi}_u}{\widehat{\psi}_l^2} n\psi - \psi\widetilde{\varsigma}_l^2 n - \frac{1-\psi}{\psi_l^2}(\psi_l\widehat{R}_{2,l}+1)^2 n\psi \\
&= \widehat{\tau}_l^2(N - n\psi) + n\psi(\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_l^2) - n\psi(1-\psi)(\widehat{R}_{2,l}^2 + 2\widehat{R}_{2,l}/\psi_l) \\
&\quad + n\psi(1-\widehat{\psi}_u)\left(\frac{1}{\widehat{\psi}_l^2} - \frac{1}{\psi_l^2}\right) - n\psi\left(\frac{\widehat{\psi}_u}{\psi_l^2} - \frac{\psi_u}{\psi_l^2}\right) + n\psi\underbrace{\left(\frac{\psi}{\psi_l^2} - \frac{\psi_u}{\psi_l^2}\right)}_{\geq 0}.
\end{aligned}
$$

We need to show that $n\psi\widetilde{\tau}_l^2 - N\widehat{\tau}_l^2$ is not too large. Hence we can use the following bound by triangle inequality and ignoring the negative terms

$$\begin{aligned}
\mathbb{E}\left(1 - \frac{\widehat{\tau}_l\sqrt{N}}{\widetilde{\tau}_l\sqrt{n\psi}}\middle|\mathbf{Z}^n\right) &\leq \mathbb{E}\left\{\frac{\left(\widetilde{\tau}_l\sqrt{n\psi} - \widehat{\tau}_l\sqrt{N}\right)\left(\widetilde{\tau}_l\sqrt{n\psi} + \widehat{\tau}_l\sqrt{N}\right)}{\widetilde{\tau}_l\sqrt{n\psi}\left(\widetilde{\tau}_l\sqrt{n\psi} + \widehat{\tau}_l\sqrt{N}\right)}\middle|\mathbf{Z}^n\right\} \\
&\leq \mathbb{E}\left(\frac{\widetilde{\tau}_l^2 n\psi - \widehat{\tau}_l^2 N}{n\psi\widetilde{\tau}_l^2}\middle|\mathbf{Z}^n\right) \\
&\leq \frac{\bar{\tau}_l^2}{n\psi\widetilde{\tau}_l^2}|N - n\psi| + \frac{\mathbb{E}\left(|\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_l^2|\middle|\mathbf{Z}^n\right)}{\widetilde{\tau}_l^2} + \frac{(1-\psi)}{\psi_l\widetilde{\tau}_l^2}\left(\psi_l\widehat{R}_{2,l}^2 + 2|\widehat{R}_{2,l}|\right) \\
&\quad + \frac{1}{\widetilde{\tau}_l^2}\mathbb{E}\left(\left|\frac{1}{\widehat{\psi}_l^2} - \frac{1}{\psi_l^2}\right|\middle|\mathbf{Z}^n\right) + \frac{1}{\widetilde{\tau}_l^2}\mathbb{E}\left(\left|\frac{\widehat{\psi}_u}{\psi_l^2} - \frac{\psi_u}{\psi_l^2}\right|\middle|\mathbf{Z}^n\right).
\end{aligned}$$

We have the following bounds

- $\mathbb{E}|N - n\psi| \leq \sqrt{n\psi(1-\psi)}$
- $\mathbb{E}|\widehat{\varsigma}_l - \widetilde{\varsigma}_l| \lesssim 1/\sqrt{N}$.

Further notice that by triangle inequality we have

$$\begin{aligned}
&\mathbb{E}\left\{(1-\widehat{\psi}_u)\left|\frac{1}{\widehat{\psi}_l^2} - \frac{1}{\psi_l^2}\right|\middle|\mathbf{Z}^n\right\} \leq \mathbb{E}\left\{\left|\frac{1}{\widehat{\psi}_l^2} - \frac{1}{\psi_l^2}\right|\middle|\mathbf{Z}^n\right\} \\
&\leq \mathbb{E}\left[\{(\mathbb{Q}_N - \mathbb{Q})\widehat{\varphi}_l\}^2\middle|\mathbf{Z}^n\right] + \widehat{R}_{2,l}^2 + 2\left(\psi_l^{-1} + |\widehat{R}_{2,l}|\right)\mathbb{E}\left\{|(\mathbb{Q}_N - \mathbb{Q})\widehat{\varphi}_l|\middle|\mathbf{Z}^n\right\} + 2\psi_l^{-1}|\widehat{R}_{2,l}| \\
&\leq \frac{\widetilde{\varsigma}_l^2}{N} + \widehat{R}_{2,l}^2 + 2\left(\psi_l^{-1} + |\widehat{R}_{2,l}|\right)\frac{\widetilde{\varsigma}_l}{\sqrt{N}} + 2\frac{|\widehat{R}_{2,l}|}{\psi_l}.
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{1}{\psi_l^2}\mathbb{E}|\widehat{\psi}_u - \psi_u| &= \mathbb{E}\left|\frac{\psi_u^{-1} - \widehat{\psi}_u^{-1}}{\widehat{\psi}_u^{-1}\psi_u^{-1}}\right| \\
&= \frac{1}{\psi_l^2}\mathbb{E}\left|\frac{(\mathbb{Q}_N - \mathbb{Q})\widehat{\varphi}_u + \widehat{R}_{2,u}}{\widehat{\psi}_u^{-1}\psi_u^{-1}}\right| \leq \frac{\widetilde{\varsigma}_u}{\psi_l\sqrt{N}} + \frac{|\widehat{R}_{2,u}|}{\psi_l},\ \text{(since } \psi_u/\psi_l,\ \widehat{\psi}_u \leq 1\text{)}.
\end{aligned}$$

Combining the above results together, we get the following

$$\begin{aligned}
\mathbb{E}\left(1 - \frac{\widehat{\tau}_l\sqrt{N}}{\widetilde{\tau}_l\sqrt{n\psi}}\middle|\mathbf{Z}^n\right) &= \frac{\bar{\tau}^2}{\widetilde{\tau}_l^2}\left|\frac{N}{n\psi} - 1\right| + \frac{\mathbb{E}|\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_l^2|}{\widetilde{\tau}_l^2} + \frac{(1-\psi)}{\psi_l\widetilde{\tau}_l^2}\left(\psi_l\widehat{R}_{2,l}^2 + 2|\widehat{R}_{2,l}|\right) \\
&+ \frac{1}{\widetilde{\tau}_l^2}\left\{\frac{\widetilde{\varsigma}_l^2}{N} + \widehat{R}_{2,l}^2 + 2\left(\psi_l^{-1} + |\widehat{R}_{2,l}|\right)\frac{\widetilde{\varsigma}_l}{\sqrt{N}} + 2\psi_l^{-1}|\widehat{R}_{2,l}| + \frac{\widetilde{\varsigma}_u}{\psi_l\sqrt{N}} + \frac{|\widehat{R}_{2,u}|}{\psi_l}\right\} \\
&\leq \frac{\bar{\tau}^2}{\widetilde{\tau}_l^2}\left|\frac{N}{n\psi} - 1\right| + \frac{\mathbb{E}|\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_l^2|}{\widetilde{\tau}_l^2} + 2|\widehat{R}_{2,l}|\frac{2 - \psi + \psi_l\widetilde{\varsigma}_l/\sqrt{N}}{\psi_l\widetilde{\tau}_l^2} \\
&+ \frac{1}{\widetilde{\tau}_l^2}\left(\frac{\widetilde{\varsigma}_l^2}{N} + \frac{2\widetilde{\varsigma}_l + \widetilde{\varsigma}_u}{\psi_l\sqrt{N}} + \frac{|\widehat{R}_{2,u}|}{\psi_l}\right) + \frac{(2-\psi)}{\widetilde{\tau}_l^2}\widehat{R}_{2,l}^2.
\end{aligned}$$

Hence, in the large sample case, we have the following simplified bound (analogously for $\widehat{\tau}_u$)

$$\mathbb{E}\left(1 - \frac{\widehat{\tau}_l\sqrt{N}}{\widetilde{\tau}_l\sqrt{n\psi}}\right) = \mathbb{E}\left\{\mathbb{E}\left(1 - \frac{\widehat{\tau}_l\sqrt{N}}{\widetilde{\tau}_l\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)\right\} \lesssim \frac{1}{\sqrt{n}} + \mathbb{E}|\widehat{R}_{2,l}|. \tag{B.4}$$

2. Similarly for the upper limit,

$$\begin{aligned}
\widehat{\tau}_u^2 N - \widetilde{\tau}_u^2 n\psi = \widehat{\tau}_u^2(N - n\psi) + n\psi(\widehat{\varsigma}_u^2 - \widetilde{\varsigma}_u^2) - n\psi(1-\psi)(\widehat{R}_{2,u}^2 + 2\widehat{R}_{2,u}/\psi_u) \\
+ n\psi\left(\frac{1}{\widehat{\psi}_u^2} - \frac{1}{\psi_u^2} - \frac{1}{\widehat{\psi}_u} + \frac{1}{\psi_u}\right) + n\psi\underbrace{\left(\frac{\psi}{\psi_u^2} - \frac{1}{\psi_u}\right)}_{\geq 0}.
\end{aligned}$$

Further notice that by triangle inequality we have

$$\begin{aligned}
&\mathbb{E}\left\{\left|\frac{1}{\widehat{\psi}_u^2} - \frac{1}{\psi_u^2} - \frac{1}{\widehat{\psi}_u} + \frac{1}{\psi_u}\right|\middle|\mathbf{Z}^n\right\} \\
&\leq \frac{\widetilde{\varsigma}_u^2}{N} + \widehat{R}_{2,u}^2 + \left|2\psi_u^{-1} + 2\widehat{R}_{2,u} - 1\right|\frac{\widetilde{\varsigma}_u}{\sqrt{N}} + (2\psi_u^{-1} - 1)|\widehat{R}_{2,u}|.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left(1 - \frac{\widehat{\tau}_u\sqrt{N}}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right) &\leq \frac{\bar{\tau}^2}{\widetilde{\tau}_u^2}\left|\frac{N}{n\psi} - 1\right| + \frac{\mathbb{E}|\widehat{\varsigma}_u^2 - \widetilde{\varsigma}_u^2|}{\widetilde{\tau}_u^2} + \frac{(1-\psi)}{\psi_u\widetilde{\tau}_u^2}\left(\psi_u\widehat{R}_{2,u}^2 + 2|\widehat{R}_{2,u}|\right) \\
&+ \frac{1}{\widetilde{\tau}_u^2}\left\{\frac{\widetilde{\varsigma}_u^2}{N} + \widehat{R}_{2,u}^2 + \left|2\psi_u^{-1} + 2\widehat{R}_{2,u} - 1\right|\frac{\widetilde{\varsigma}_u}{\sqrt{N}} + |2\psi_u^{-1} - 1||\widehat{R}_{2,u}|\right\} \\
&\leq \frac{\bar{\tau}^2}{\widetilde{\tau}_u^2}\left|\frac{N}{n\psi} - 1\right| + \frac{\mathbb{E}|\widehat{\varsigma}_u^2 - \widetilde{\varsigma}_u^2|}{\widetilde{\tau}_u^2} + \frac{2|\widehat{R}_{2,l}|}{\psi_u\widetilde{\tau}_u^2}\left(1 - \psi + |1 - \psi_u/2| + \frac{\psi_u\widetilde{\varsigma}_u}{\sqrt{N}}\right) \\
&+ \frac{1}{\widetilde{\tau}_u^2}\left(\frac{\widetilde{\varsigma}_u^2}{N} + \frac{|2\psi_u^{-1} - 1|}{\sqrt{N}}\widetilde{\varsigma}_u\right) + \frac{\widehat{R}_{2,u}^2(2-\psi)}{\widetilde{\tau}_u^2}.
\end{aligned}$$

Hence, in the large sample case, we have the following simplified bound (analogously for $\widehat{\tau}_u$)

$$\mathbb{E}\left(1 - \frac{\widehat{\tau}_u\sqrt{N}}{\widetilde{\tau}_u\sqrt{n\psi}}\right) = \mathbb{E}\left\{\mathbb{E}\left(1 - \frac{\widehat{\tau}_u\sqrt{N}}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)\right\} \lesssim \frac{1}{\sqrt{n}} + \mathbb{E}|\widehat{R}_{2,u}|. \tag{B.5}$$

3. For the last difference, we follow similarly as we did for the first one.

$$\begin{aligned}\widehat{\tau}_l^2 N \ - \widetilde{\tau}_u^2 n\psi &= \widehat{\tau}_l^2(N - n\psi) + n\psi(\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_u^2) - n\psi(1-\psi)\left(\widehat{R}_{2,u}^2 + 2\frac{|\widehat{R}_{2,u}|}{\psi_u} + \Delta^2 + 2\frac{\Delta}{\psi_l}\right) \\ &+ n\psi(1-\widehat{\psi}_u)\left(\frac{1}{\widehat{\psi}_l^2} - \frac{1}{\psi_l^2}\right) - n\psi\left(\frac{\widehat{\psi}_u}{\psi_l^2} - \frac{\psi_u}{\psi_l^2}\right) + n\psi\underbrace{\left(\frac{\psi}{\psi_l^2} - \frac{\psi_u}{\psi_l^2}\right)}_{\geq 0}\left(\text{since } \frac{1}{\psi_u} = \Delta + \frac{1}{\psi_l}\right).\end{aligned}$$

The expected difference is bounded as follows

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{\widehat{\tau}_l\sqrt{N}} - \frac{1}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right) \leq \mathbb{E}\left(\frac{\widetilde{\tau}_u^2 n\psi - \widehat{\tau}_l^2 N}{\widehat{\tau}_l\widetilde{\tau}_u^2\sqrt{N}n\psi}\middle|\mathbf{Z}^n\right) \\ &\leq \frac{\bar{\tau}}{\widetilde{\tau}_u^2\sqrt{N}}\left|\frac{N}{n\psi} - 1\right| + \frac{\mathbb{E}|\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_u^2|}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}} + \frac{(1-\psi)}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}}\left(\widehat{R}_{2,u}^2 + 2\frac{|\widehat{R}_{2,u}|}{\psi_u} + \Delta^2 + 2\frac{\Delta}{\psi_l}\right) \\ &+ \frac{1}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}}\left\{\frac{\widetilde{\varsigma}_l^2}{N} + \widehat{R}_{2,l}^2 + 2\left(\psi_l^{-1} + |\widehat{R}_{2,l}|\right)\frac{\widetilde{\varsigma}_l}{\sqrt{N}} + 2\psi_l^{-1}|\widehat{R}_{2,l}| + \frac{\widetilde{\varsigma}_u}{\psi_l\sqrt{N}} + \frac{|\widehat{R}_{2,u}|}{\psi_l}\right\} \\ &\leq \frac{\bar{\tau}}{\widetilde{\tau}_u^2\sqrt{N}}\left|\frac{N}{n\psi} - 1\right| + \frac{\mathbb{E}|\widehat{\varsigma}_l^2 - \widetilde{\varsigma}_u^2|}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}} + \frac{|\widehat{R}_{2,u}|}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}}\left(\frac{2-2\psi}{\psi_u} + \frac{1}{\psi_l}\right) + \frac{2|\widehat{R}_{2,l}|}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}}\left(\frac{\widetilde{\varsigma}_l}{\sqrt{N}} + \frac{1}{\psi_l}\right) \\ &+ \frac{1}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}}\left\{\frac{\widetilde{\varsigma}_l^2}{N} + \frac{2\widetilde{\varsigma}_l + \widetilde{\varsigma}_u}{\psi_l\sqrt{N}} + \widehat{R}_{2,u}^2(1-\psi) + \widehat{R}_{2,l}^2\right\} + \frac{(1-\psi)}{\widetilde{\tau}_u^2\underline{\tau}\sqrt{N}}\left(\Delta^2 + 2\frac{\Delta}{\psi_l}\right).\end{aligned}$$

For the large sample case,

$$\mathbb{E}\left\{\mathbb{E}\left(\frac{1}{\widehat{\tau}_l\sqrt{N}} - \frac{1}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)\right\} = \mathbb{E}\left(\frac{1}{\widehat{\tau}_l\sqrt{N}} - \frac{1}{\widetilde{\tau}_u\sqrt{n\psi}}\right) \lesssim \frac{1}{\sqrt{n}} + \frac{\mathbb{E}|\widehat{R}_{2,l}| + \mathbb{E}|\widehat{R}_{2,u}|}{\sqrt{n}}. \tag{B.6}$$

Proof for the first difference B.1

For the first difference

$$\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{n}_l - c_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + c_N\sqrt{N}\widehat{\tau}_u\right)\\
&= (1-\alpha) - \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\sqrt{N}\widehat{\tau}_l} \leq \frac{n_u - n_l}{\sqrt{N}\widehat{\tau}_l} + c_N\right) + \mathbb{P}\left(\frac{\widehat{n}_u - n_u}{\sqrt{N}\widehat{\tau}_u} \leq -c_N\right)\\
&\leq \Phi\left(c_N + \frac{n_u - n_l}{\sqrt{N}\widehat{\tau}_l}\right) - \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\sqrt{N}\widehat{\tau}_l} \leq c_N + \frac{n_u - n_l}{\sqrt{N}\widehat{\tau}_l}\right)\\
&+ \mathbb{P}\left(\frac{\widehat{n}_u - n_u}{\sqrt{N}\widehat{\tau}_u} \leq -c_N\right) - \Phi\left(-c_N\right)\\
&+ \Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) - \Phi\left(c_N + \frac{n_u - n_l}{\sqrt{N}\widehat{\tau}_l}\right)\\
&\text{since } \Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) - \Phi(-c_N) = 1-\alpha.
\end{aligned}$$

The first two terms are bounded above by Berry-Esseen and for the third term, we can use mean value theorem as follows,

$$\Phi\left(c_N + \frac{n_u - n_l}{\sqrt{N}\widehat{\tau}_l}\right) - \Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) = n\psi\Delta\phi(t_3)\left(\frac{1}{\sqrt{N}\widehat{\tau}_l} - \frac{1}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right),$$

for some t_3 between $c_N + \frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l}$ and $c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}$, where $n_u - n_l = n\psi\Delta$. We can show that this term is not too positive using results from bounds B.4 and B.6.

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{\widehat{\tau}_l\sqrt{N}} - \frac{1}{(\widetilde{\tau}_l \vee \widetilde{\tau}_u)\sqrt{n\psi}}\middle|\mathbf{Z}^n\right\} &\leq \mathbb{E}\left(\frac{1}{\widehat{\tau}_l\sqrt{N}} - \frac{1}{\widetilde{\tau}_l\sqrt{n\psi}}\middle|\mathbf{Z}^n\right) \vee \mathbb{E}\left(\frac{1}{\widehat{\tau}_l\sqrt{N}} - \frac{1}{\widetilde{\tau}_u\sqrt{n\psi}}\middle|\mathbf{Z}^n\right)\\
&\lesssim \frac{1}{\sqrt{N}} + \frac{|\widehat{R}_{2,l}| + |\widehat{R}_{2,u}|}{\sqrt{N}}.
\end{aligned}$$

Since $c_N > 0$, we can bound the $\phi(t_3)$ term as follows

$$\begin{aligned}
n\psi\Delta\phi(t_3) \leq n\psi\Delta\phi\left(\frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l \vee \sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l}\right) &= \frac{n\psi\Delta}{\sqrt{2\pi}} exp\left\{-\frac{n^2\psi^2\Delta^2}{2(\sqrt{N}\widehat{\tau}_l \vee \sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l)^2}\right\}\\
&\leq \sqrt{\frac{2}{\pi}}\frac{(\sqrt{N}\widehat{\tau}_l \vee \sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l)^2}{n\psi\Delta},
\end{aligned}$$

since $e^{-w} \leq 1/w \ \forall \ w > 0$.

Thus,

$$\begin{aligned}
&\mathbb{E}\left\{\Phi\left(c_N + \frac{n_u - n_l}{\sqrt{N}\widehat{\tau}_l}\right) - \Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right)\right\}\\
&\leq \mathbb{E}\left\{\left(\frac{1}{\sqrt{N}\widehat{\tau}_l} - \frac{1}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right)\sqrt{\frac{2}{\pi}}\frac{(\sqrt{N}\widehat{\tau}_l \vee \sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l)^2}{n\psi\Delta}\right\}.
\end{aligned} \tag{B.7}$$

By Berry-Esseen, we have for any t' that

$$\Phi(t') - \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3} \leq \mathbb{P}\left(\frac{\widehat{n}_l - n_l - n\psi\widehat{R}_{2,l}}{\sqrt{n\psi}\widetilde{\tau}_l} \leq t' \;\middle|\; \mathbf{Z}^n\right) \leq \Phi(t') + \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3},$$

where $\rho_l = \mathbb{E}\left[|\mathbb{1}(\mathbf{Y} \neq \mathbf{0})(\widehat{\phi}_l - \mathbb{Q}\widehat{\phi}_l) + \{\mathbb{1}(\mathbf{Y} \neq \mathbf{0}) - \psi\}(\widehat{R}_{2,l} + \psi_l^{-1})|^3 \middle| \mathbf{Z}^n\right]$.
Taking $t' = \frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}t - \frac{n\psi\widehat{R}_{2,l}}{\sqrt{n\psi}\widetilde{\tau}_l}$, this implies

$$\begin{aligned}\Phi\left(\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}t - \frac{\sqrt{n\psi}\widehat{R}_{2,l}}{\widetilde{\tau}_l}\right) - \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3} &\leq \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\sqrt{N}\widehat{\tau}_l} \leq t \;\middle|\; \mathbf{Z}^n\right) \\ &\leq \Phi\left(\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}t - \frac{\sqrt{n\psi}\widehat{R}_{2,l}}{\widetilde{\tau}_l}\right) + \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3}.\end{aligned}$$

Therefore, by the mean value theorem and the fact that $\phi(w) \leq \frac{1}{\sqrt{2\pi}}\,\forall\, w$

$$\begin{aligned}\Phi(t) - \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\sqrt{N}\widehat{\tau}_l} \leq t \;\middle|\; \mathbf{Z}^n\right) &\leq\; \Phi(t) - \Phi\left(t\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l} - \frac{\sqrt{n\psi}\widehat{R}_{2,l}}{\widetilde{\tau}_l}\right) + \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3} \\ &\leq\; \phi(\theta)\left\{\left(1 - \frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)t + \frac{\sqrt{n\psi}|\widehat{R}_{2,l}|}{\widetilde{\tau}_l}\right\} + \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3},\end{aligned}$$

for some θ between t and $t\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l} - \frac{\sqrt{n\psi}\widehat{R}_{2,l}}{\widetilde{\tau}_l}$.

Further by substituting $t = c_N + \frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l}$, we see that $\theta \geq \frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l} \wedge \frac{\sqrt{n\psi}|\Delta - \widehat{R}_{2,l}|}{\widetilde{\tau}_l} = \frac{n\psi(\Delta \wedge |\Delta - \widehat{R}_{2,l}|)}{(\sqrt{N}\widehat{\tau}_l) \vee (\sqrt{n\psi}\widetilde{\tau}_l)}$. Thus, using the inequality $e^{-w} \leq 1/w \;\forall\; w > 0$, we get

$$\phi(\theta) \leq \sqrt{\frac{2}{\pi}}\frac{(N\widehat{\tau}_l^2) \vee (n\psi\widetilde{\tau}_l^2)}{n^2\psi^2(\Delta \wedge |\Delta - \widehat{R}_{2,l}|)^2}, \frac{1}{\sqrt{2\pi}}.$$

Substituting these bounds in the above inequality, we get the following.

$$\begin{aligned}\Phi(t) - \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\sqrt{N}\widehat{\tau}_l} \leq t \;\middle|\; \mathbf{Z}^n\right) &\leq\; \Phi(t) - \Phi\left(t\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l} - \frac{\sqrt{n\psi}\widehat{R}_{2,l}}{\widetilde{\tau}_l}\right) + \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3} \\ &\leq\; \sqrt{\frac{2}{\pi}}\frac{(N\widehat{\tau}_l^2) \vee (n\psi\widetilde{\tau}_l^2)}{n\psi(\Delta \wedge |\Delta - \widehat{R}_{2,l}|)^2}\left(1 - \frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{\Delta}{\sqrt{N}\widehat{\tau}_l} \\ &\quad + \left(1 - \frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{c_N}{\sqrt{2\pi}} + \frac{\sqrt{n\psi}|\widehat{R}_{2,l}|}{\sqrt{2\pi}\widetilde{\tau}_l} + \frac{C\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3},\end{aligned} \tag{B.8}$$

A similar result follows for $\widehat{n}_u$.

$$\begin{aligned}\mathbb{P}\left(\frac{\widehat{n}_u - n_u}{\sqrt{N}\widehat{\tau}_u} \leq -t \,\middle|\, \mathbf{Z}^n\right) - \Phi(-t) \leq\ & \Phi\left(-t\frac{\sqrt{N}\widehat{\tau}_u}{\sqrt{n\psi}\widetilde{\tau}_u} - \frac{\sqrt{n\psi}\widehat{R}_{2,u}}{\widetilde{\tau}_u}\right) - \Phi(-t) + \frac{C\rho_u}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_u^3} \\ \leq\ & \frac{1}{\sqrt{2\pi}}\left\{\left(1 - \frac{\sqrt{N}\widehat{\tau}_u}{\sqrt{n\psi}\widetilde{\tau}_u}\right)t + \frac{\sqrt{n\psi}|\widehat{R}_{2,u}|}{\widetilde{\tau}_u}\right\} + C\left(\frac{\rho_u}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_u^3}\right).\end{aligned} \tag{B.9}$$

Proof for the second bound B.2

The second difference can be re-written as

$$\begin{aligned}&\mathbb{P}(\widetilde{C}_N < c_N)\mathbb{P}\left(\widehat{n}_l - c_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \mid \widetilde{C}_N < c_N\right) \\ &+ \mathbb{P}(\widetilde{C}_N < c_N)\mathbb{P}\left(\widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u \leq n_u \leq \widehat{n}_u + c_N\sqrt{N}\widehat{\tau}_u \mid \widetilde{C}_N < c_N\right) \\ &- \mathbb{P}(\widetilde{C}_N > c_N)\mathbb{P}\left(\widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_l - c_N\sqrt{N}\widehat{\tau}_l \mid \widetilde{C}_N > c_N\right) \\ &- \mathbb{P}(\widetilde{C}_N > c_N)\mathbb{P}\left(\widehat{n}_u + c_N\sqrt{N}\widehat{\tau}_u \leq n_u \leq \widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u \mid \widetilde{C}_N > c_N\right).\end{aligned}$$

The probabilities are bounded above by 1. Assuming that $\widehat{\psi}$, $\sqrt{N}\widehat{\tau}_l$ and $\sqrt{N}\widehat{\tau}_u$ have continuous densities (to avoid high mass in a small area), it is sufficient to show that the positive terms are not too positive. This difference is bounded above by

$$\begin{aligned}&2\mathbb{P}(c_N - \widetilde{C}_N > \eta) + 2\mathbb{P}(0 < c_N - \widetilde{C}_N \leq \eta) \\ \leq\ & 2\mathbb{P}(c_N - \widetilde{C}_N > \eta) + 2\eta\theta,\end{aligned} \tag{B.10}$$

where θ is the maximum value of the density of $(\widehat{n}_u - n_u)/\sqrt{N}\widehat{\tau}_u$ and $(\widehat{n}_l - n_u)/\sqrt{N}\widehat{\tau}_l$.

We will show the following is bounded above.

$$\mathbb{P}(c_N - \widetilde{C}_N > \eta)$$

To prove the same, we will show that the following is bounded for any given $\eta > 0$.

$$\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right).$$

This however requires as additional assumption. We use the following finite sample assumption modifying the assumption in Imbens and Manski (2004) (assumption 1 (iii)).

Assumption 8. *For a given $\epsilon > 0$ and a constant c, there exists N_0 and $\upsilon > 0$ such that for all $N > N_0$*

$$\mathbb{P}\left(|\widehat{\Delta} - \Delta| > cn^{-\upsilon}\right) < \epsilon.$$

Notice that we can break the event above into the following three terms.

$$\begin{aligned}
&\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta \\
&= \mathbb{1}(N\widehat{\Delta} \leq n\psi\Delta) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right\} \\
&+ \mathbb{1}(N\widehat{\Delta} > n\psi\Delta, |\widehat{\Delta} - \Delta| \leq cn^{-\upsilon}) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right\} \\
&+ \mathbb{1}(N\widehat{\Delta} > n\psi\Delta, |\widehat{\Delta} - \Delta| > cn^{-\upsilon}) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right\}.
\end{aligned}$$

The goal is to show that the probability of this event is not too large. So we start by maximizing this probability and show that it is bounded. The first term has zero probability and hence, can be dropped. For the second term, notice the following

$$\begin{aligned}
&\mathbb{1}\left\{N\widehat{\Delta} > n\psi\Delta, |\widehat{\Delta} - \Delta| \leq cn^{-\upsilon}, \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right\} \\
&\leq \mathbb{1}\left\{|\widehat{\Delta} - \Delta| \leq cn^{-\upsilon}, \phi\left(\frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) \times \frac{\widehat{\Delta}|N - n\psi| + n\psi|\widehat{\Delta} - \Delta|}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)} > \eta\right\},
\end{aligned}$$

where the second bound follows by mean value theorem and triangle inequality.Now, by Markov's inequality we get the following bound on the second term.

$$\mathbb{E}\left\{\frac{1}{\eta} \times \phi\left(\frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) \times \frac{|N - n\psi|\widehat{\Delta} + n\psi cn^{-\upsilon}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right\} = \mathbb{E}\left[\frac{|N - n\psi|\widehat{\Delta} + \psi cn^{1-\upsilon}}{\eta\sqrt{n\psi}\sqrt{2\pi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\ exp\left\{-\frac{n\psi\Delta^2}{2(\widetilde{\tau}_u \vee \widetilde{\tau}_l)^2}\right\}\right].$$

For the third term, it is easy to see that it is bounded above by

$$\mathbb{P}(|\widehat{\Delta} - \Delta| > cn^{-\upsilon}) < \frac{1}{\sqrt{n}},$$

when N is sufficiently large and c is a constant chosen appropriately.

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right)\\
&\leq \mathbb{E}\left[\frac{|N-n\psi|\widehat{\Delta} + \psi c n^{1-\upsilon}}{\eta\sqrt{n\psi}\sqrt{2\pi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\ exp\left\{-\frac{n\psi\Delta^2}{2(\widetilde{\tau}_u \vee \widetilde{\tau}_l)^2}\right\}\right] + \frac{1}{\sqrt{n}}.
\end{aligned}$$

Next, to show that $P(c_N - \widetilde{C}_N > \eta)$ is bounded above, we use mean value theorem. We will use the previously proved result and show that it is equivalent to $\mathbb{P}(c_N - \widetilde{C}_N > \eta)$.

$$\begin{aligned}
&\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right)\\
&= \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi(-\widetilde{C}_N) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) + \Phi(-\widetilde{C}_N)\\
&= \Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) - \Phi(-c_N) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) + \Phi(-\widetilde{C}_N).
\end{aligned}$$

The third equality follows from the definitions of $\widetilde{C}_N$ and c_N. By mean value theorem,

$$\begin{aligned}
&\Phi\left(c_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) - \Phi(-c_N) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) + \Phi(-\widetilde{C}_N)\\
&= \phi(t_1)(c_N - \widetilde{C}_N) + \phi(t_2)(c_N - \widetilde{C}_N), \text{for some numbers } t_1 \text{ and } t_2.
\end{aligned}$$

Also, if $\sqrt{n\psi}\Delta/(\widetilde{\tau}_u \vee \widetilde{\tau}_l)$, $\widetilde{C}_N$ and c_N are bounded above, then $\phi(t_1) + \phi(t_2)$ is bounded away from zero. Note that $\phi(t_1) + \phi(t_2) > \phi(z_{\alpha/2}) = \alpha/2$ since c_N, $\widetilde{C}_N \leq z_{\alpha/2}$. Hence, we have the following equivalence for any given $\eta > 0$.

$$\mathbb{P}\left(c_N - \widetilde{C}_N > \frac{\eta}{\phi(t_1) + \phi(t_2)}\right) = \mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) - \Phi\left(\widetilde{C}_N + \frac{\sqrt{n\psi}\Delta}{\widetilde{\tau}_u \vee \widetilde{\tau}_l}\right) > \eta\right).$$

Thus, the second difference B.2 is bounded above by

$$2\mathbb{P}\left(c_N - \widetilde{C}_N > \eta\right) + 2\eta\ \theta.$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{n}_l - c_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + c_N\sqrt{N}\widehat{\tau}_u\right) - \mathbb{P}\left(\widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u\right)\\
&\leq 2\mathbb{E}\left[\frac{|N-n\psi|\widehat{\Delta} + \psi c n^{1-\upsilon}}{\eta(\alpha/2)\sqrt{n\psi}\sqrt{2\pi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\ exp\left\{-\frac{n\psi\Delta^2}{2(\widetilde{\tau}_u \vee \widetilde{\tau}_l)^2}\right\}\right] + \frac{2}{\sqrt{n}} + 2\eta\ \theta
\end{aligned} \tag{B.11}$$

Proof for the third difference B.3

The third difference can be re-written as

$$\begin{aligned}
&\mathbb{P}(\bar{C}_N < \widetilde{C}_N)\mathbb{P}\left(\widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \mid \bar{C}_N < \widetilde{C}_N\right)\\
+\ &\mathbb{P}(\bar{C}_N < \widetilde{C}_N)\mathbb{P}\left(\widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u \leq n_u \leq \widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u \mid \bar{C}_N < \widetilde{C}_N\right)\\
-\ &\mathbb{P}(\bar{C}_N > \widetilde{C}_N)\mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \leq n_u \leq \widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \mid \bar{C}_N > \widetilde{C}_N\right)\\
-\ &\mathbb{P}(\bar{C}_N > \widetilde{C}_N)\mathbb{P}\left(\widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u \leq n_u \leq \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u \mid \bar{C}_N > \widetilde{C}_N\right).
\end{aligned}$$

We will approach in a similar way as we did for the second difference B.11. The probabilities are bounded above by 1. We just need to show that this difference is not too positive. Assuming that $\widehat{\psi}$ and $\sqrt{N}\widehat{\tau}$ have uniformly continuous densities (to avoid high mass in a small area), it is sufficient to show the following

$$\mathbb{P}(\widetilde{C}_N - \bar{C}_N > \eta)$$

is bounded above.

To prove the same, we will show that the following

$$\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\right)$$

is bounded for a given $\eta > 0$.

Notice that we can break the event above into the following terms.

$$\begin{aligned}
&\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\\
&=\ \mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} \leq 1\right) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\right\}\\
&+\ \mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} > 1\right) \times \mathbb{1}\left\{\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\right\}.
\end{aligned}$$

The goal is to show that the probability of this event is not too large. So we start by maximizing this probability and show that it is bounded. The first term has zero probability and hence, can be safely dropped. For the second term, notice the following

$$\begin{aligned}
&\mathbb{1}\left\{\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} > 1,\, \Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\right\} \\
&\leq\ \mathbb{1}\left\{\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} > 1,\, \phi\left(\frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) \times \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\left(1 - \frac{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\right\},
\end{aligned}$$

where the second bound follows by mean value theorem and the properties of the normal density ϕ. Now, by Markov's inequality we get the following bound for the expectation of the above term

$$\mathbb{E}\left[\frac{1}{\eta}\phi\left(\frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) \times \frac{\sqrt{N}\widehat{\Delta}\ \{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l) - \sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)\}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)(\widehat{\tau}_u \vee \widehat{\tau}_l)} \times \mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} > 1\right)\right],$$

$$\text{since } \mathbb{E}\{(A - C)\mathbb{1}(A - C > 0)\} < \mathbb{E}\{A\mathbb{1}(A - C > 0)\} \leq \mathbb{E}|A|,\ \text{ for variables } A \,\&\, C.$$

This quantity can be bounded above using the results of B.4 and B.5. Also, notice that

$$\frac{\sqrt{n\psi}\widetilde{\tau}_u \vee \sqrt{n\psi}\widetilde{\tau}_l - \sqrt{N}\widehat{\tau}_u \vee \sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)} \leq \frac{\sqrt{n\psi}\widetilde{\tau}_l - \sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)} \vee \frac{\sqrt{n\psi}\widetilde{\tau}_u - \sqrt{N}\widehat{\tau}_u}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}.$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \eta\right) \\
&\quad \leq \mathbb{E}\left[\frac{\sqrt{N}\widehat{\Delta}\ \{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l) - \sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)\}}{\eta\sqrt{2\pi n\psi}(\widehat{\tau}_u \vee \widehat{\tau}_l)(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\ exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\tau}_u \vee \widehat{\tau}_l)^2}\right\}\right].
\end{aligned}$$

Next to show that $\mathbb{P}(\widetilde{C}_N - \bar{C}_N > \eta)$ is bounded above, we can use mean value theorem similar to what we did for the second difference.

Hence, we have the following for any given $\eta > 0$

$$\mathbb{P}\left(\widetilde{C}_N - \bar{C}_N > \eta\right) \leq \mathbb{P}\left(\Phi\left(\widetilde{C}_N + \frac{\sqrt{N}\widehat{\Delta}}{\widehat{\tau}_u \vee \widehat{\tau}_l}\right) - \Phi\left(\widetilde{C}_N + \frac{N\widehat{\Delta}}{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\right) > \frac{\eta\alpha}{2}\right).$$

Thus, the third difference B.3 is bounded above by

$$2\mathbb{P}\left(\widetilde{C}_N - \bar{C}_N > \eta\right) + 2\eta\ \theta,$$

where θ is the maximum value of the density of $(\widehat{n}_u - n_u)/\sqrt{N}\widehat{\tau}_u$ and $(\widehat{n}_l - n_u)/\sqrt{N}\widehat{\tau}_l$.

$$2\mathbb{E}\left\{\mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \le n_u \le \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u\right) - \mathbb{P}\left(\widehat{n}_l - \widetilde{C}_N\sqrt{N}\widehat{\tau}_l \le n_u \le \widehat{n}_u + \widetilde{C}_N\sqrt{N}\widehat{\tau}_u\right)\right\}$$
$$\le \mathbb{E}\left[\frac{\sqrt{N}\widehat{\Delta}\ \{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l) - \sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)\}}{\eta(\alpha/2)\sqrt{2\pi n\psi}(\widehat{\tau}_u \vee \widehat{\tau}_l)(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}\ exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\tau}_u \vee \widehat{\tau}_l)^2}\right\} \times \mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u \vee \widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u \vee \widehat{\tau}_l)} > 1\right)\right] + 2\eta\ \theta. \tag{B.12}$$

Combining the bounds

Now, we can show that the estimated confidence internal contains n_u with probability $1-\alpha$ and some error term that is not too negative. Similarly, one can show for n_l. And hence, the result follows for the target parameter ψ.

$$\begin{aligned}
&(1-\alpha) - \mathbb{P}\left(\widehat{n}_l - \bar{C}_N\sqrt{N}\widehat{\tau}_l \le n_u \le \widehat{n}_u + \bar{C}_N\sqrt{N}\widehat{\tau}_u\right) \\
\le\ &\mathbb{E}\left\{\Phi\left(c_N + \frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l}\right) - \mathbb{P}\left(\frac{\widehat{n}_l - n_l}{\sqrt{N}\widehat{\tau}_l} \le c_N + \frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l}\right)\right\} \\
&+\ \mathbb{E}\left\{\mathbb{P}\left(\frac{\widehat{n}_u - n_u}{\sqrt{N}\widehat{\tau}_u} \le -c_N\right) - \Phi\left(-c_N\right)\right\} \\
&+\ \mathbb{E}\left\{\Phi\left(c_N + \frac{n\psi\Delta}{\sqrt{n\psi}(\widetilde{\tau}_l \vee \widetilde{\tau}_u)}\right) - \Phi\left(c_N + \frac{n\psi\Delta}{\sqrt{N}\widehat{\tau}_l}\right)\right\} \\
&+\ 2\mathbb{P}\left(c_N - \widetilde{C}_N > \eta\right) + 2\eta\ \theta + 2\mathbb{P}\left(\widetilde{C}_N - \bar{C}_N > \eta\right) + 2\eta\ \theta.
\end{aligned}$$

Now, if $\Delta = 0$, then in the context of this paper, this condition indicates that $\widehat{\Delta} = 0$. Thus, $\bar{C}_N = \widetilde{C}_N = c_N = z_{1-\alpha/2}$. Thus, the third term above along with the second B.2 and the third differences B.3 are zero when $\Delta = 0$. To incorporate this property here, we will use indicator terms. Next, we use the results from B.7, B.8, B.9, B.11 and B.12 to obtain the following bound.

$$
\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{n\psi}|\widehat{R}_{2,l}|}{\widetilde{\tau}_l}\right) + \frac{c_N}{\sqrt{2\pi}}\mathbb{E}\left(1-\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right) + C\mathbb{E}\left(\frac{\rho_l}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_l^3}\right)\\
&+\ \sqrt{\frac{2}{\pi}}\mathbb{E}\left\{\frac{(N\widehat{\tau}_l^2)\vee(n\psi\widetilde{\tau}_l^2)}{n\psi(\Delta\wedge|\Delta-\widehat{R}_{2,l}|)^2}\left(1-\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{\Delta}{\sqrt{N}\widehat{\tau}_l}\right\}\\
&+\ \frac{1}{\sqrt{2\pi}}\mathbb{E}\left(\frac{\sqrt{n\psi}|\widehat{R}_{2,u}|}{\widetilde{\tau}_u}\right) + \frac{c_N}{\sqrt{2\pi}}\mathbb{E}\left(1-\frac{\sqrt{N}\widehat{\tau}_u}{\sqrt{n\psi}\widetilde{\tau}_u}\right) + C\mathbb{E}\left(\frac{\rho_u}{\sqrt{n}\psi^{1.5}\widetilde{\tau}_u^3}\right)\\
&+\mathbb{1}(\Delta=0)\mathbb{E}\left\{\left(\frac{1}{\sqrt{N}\widehat{\tau}_l}-\frac{1}{\sqrt{n\psi}\widetilde{\tau}_u\vee\sqrt{n\psi}\widetilde{\tau}_l}\right)\sqrt{\frac{2}{\pi}}\frac{(\sqrt{N}\widehat{\tau}_l\vee\sqrt{n\psi}\widetilde{\tau}_u\vee\sqrt{n\psi}\widetilde{\tau}_l)^2}{n\psi\Delta}\right\}\\
&+\ 2\mathbb{1}(\Delta\neq 0)\mathbb{E}\left[\frac{|N-n\psi|\widehat{\Delta}+\psi c n^{1-\upsilon}}{\eta(\alpha/2)\sqrt{n\psi}\sqrt{2\pi}(\widetilde{\tau}_u\vee\widetilde{\tau}_l)}\ exp\left\{-\frac{n\psi\Delta^2}{2(\widetilde{\tau}_u\vee\widetilde{\tau}_l)^2}\right\}+\frac{1}{\sqrt{n}}\right]+\mathbb{1}(\Delta\neq 0)4\eta\ \theta\\
&+\ 2\mathbb{1}(\Delta\neq 0)\mathbb{E}\left[\frac{\sqrt{N}\widehat{\Delta}\ \{\sqrt{n\psi}(\widetilde{\tau}_u\vee\widetilde{\tau}_l)-\sqrt{N}(\widehat{\tau}_u\vee\widehat{\tau}_l)\}}{\eta(\alpha/2)\sqrt{2\pi n\psi}(\widehat{\tau}_u\vee\widehat{\tau}_l)(\widetilde{\tau}_u\vee\widetilde{\tau}_l)}exp\left\{-\frac{N\widehat{\Delta}^2}{2(\widehat{\tau}_u\vee\widehat{\tau}_l)^2}\right\}\mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u\vee\widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u\vee\widehat{\tau}_l)}>1\right)\right].
\end{aligned}
$$

If $\eta = n^{-\kappa}$ for some $\kappa > 0$, to simplify the bound, let $\kappa = 1/2$. We choose $\eta = n^{-1/2}$ since the rate is not faster than $1/\sqrt{n}$ as can be seen from the Berry-Esseen bounds. Also, by definition $\upsilon > 0$.

Further, to simplify the exponential terms, we use the inequality $e^{-\frac{w}{2}} < \sqrt{3!/w^3}\ \forall\ w > 0$, and some rearrangement to obtain the following simplified form.

$$
\begin{aligned}
&(1-\alpha)-\mathbb{E}\left\{\mathbb{P}\left(\widehat{n}_l-\bar{C}_N\sqrt{N}\widehat{\tau}_l\leq n_u\leq\widehat{n}_u+\bar{C}_N\sqrt{N}\widehat{\tau}_u\right)\right\}\\
\leq\ &\frac{\sqrt{n\psi}}{\sqrt{2\pi}}\mathbb{E}\left(\frac{|\widehat{R}_{2,l}|}{\widetilde{\tau}_l}+\frac{|\widehat{R}_{2,u}|}{\widetilde{\tau}_u}\right)+\frac{C}{\sqrt{n}\psi^{1.5}}\mathbb{E}\left(\frac{\rho_l}{\widetilde{\tau}_l^3}+\frac{\rho_u}{\widetilde{\tau}_u^3}\right)+\mathbb{1}(\Delta\neq 0)\mathbb{E}\left\{\frac{2}{\sqrt{n}}\left(1+2\theta\right)\right\}\\
&+\ \frac{c_N}{\sqrt{2\pi}}\left\{\mathbb{E}\left(1-\frac{\sqrt{N}\widehat{\tau}_u}{\sqrt{n\psi}\widetilde{\tau}_u}\right)+\mathbb{E}\left(1-\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\right\}\\
&+\ \sqrt{\frac{2}{\pi}}\mathbb{E}\left\{\frac{(N\widehat{\tau}_l^2)\vee(n\psi\widetilde{\tau}_l^2)}{n\psi(\Delta\wedge|\Delta-\widehat{R}_{2,l}|)^2}\left(1-\frac{\sqrt{N}\widehat{\tau}_l}{\sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{\Delta}{\sqrt{N}\widehat{\tau}_l}\right\}\\
&+\ \mathbb{1}(\Delta\neq 0)\ \mathbb{E}\left\{\left(\frac{1}{\sqrt{N}\widehat{\tau}_l}-\frac{1}{\sqrt{n\psi}\widetilde{\tau}_u\vee\sqrt{n\psi}\widetilde{\tau}_l}\right)\frac{\sqrt{2}(\sqrt{N}\widehat{\tau}_l\vee\sqrt{n\psi}\widetilde{\tau}_u\vee\sqrt{n\psi}\widetilde{\tau}_l)^2}{\sqrt{\pi}n\psi\Delta}\right\}\\
&+\ \mathbb{1}(\Delta\neq 0)\frac{2\sqrt{3}}{\alpha\sqrt{\pi}\psi^2 n^{1.5}}\mathbb{E}\left[\frac{|N-n\psi|\widehat{\Delta}+\psi c n^{1-\upsilon}}{}\ \frac{(\widetilde{\tau}_u\vee\widetilde{\tau}_l)^2}{\Delta^3}\right]\\
&+\ \mathbb{1}(\Delta\neq 0)\frac{2\sqrt{3}}{\alpha\sqrt{\pi\psi}}\mathbb{E}\left[\frac{\{\sqrt{n\psi}(\widetilde{\tau}_u\vee\widetilde{\tau}_l)-\sqrt{N}(\widehat{\tau}_u\vee\widehat{\tau}_l)\}}{(\widetilde{\tau}_u\vee\widetilde{\tau}_l)}\ \frac{(\widehat{\tau}_u\vee\widehat{\tau}_l)^2}{N\widehat{\Delta}^2}\times\mathbb{1}\left(\frac{\sqrt{n\psi}(\widetilde{\tau}_u\vee\widetilde{\tau}_l)}{\sqrt{N}(\widehat{\tau}_u\vee\widehat{\tau}_l)}>1\right)\right].
\end{aligned}
$$

Now, using the upper bounds on the variance differences, we get the large sample bound as follows.

$$\frac{1}{\sqrt{n}}+\sqrt{n}\mathbb{E}(|\widehat{R}_{2,l}|+|\widehat{R}_{2,u}|).$$

□