DISSERTATION

TOPICS IN ESTIMATION FOR MESSY SURVEYS: IMPERFECT MATCHING AND NONPROBABILITY SAMPLING

Submitted by

Chien-Min Huang

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2022

Doctoral Committee:

Advisor: F. Jay Breidt

Haonan Wang
Joshua Keller
Sangmi Pallickara

ABSTRACT

TOPICS IN ESTIMATION FOR MESSY SURVEYS: IMPERFECT MATCHING AND NONPROBABILITY SAMPLING

Two problems in estimation for "messy" surveys are addressed, both requiring the combination of survey data with other data sources. The first estimation problem involves the combination of survey data with auxiliary data, when the matching of the two sources is imperfect. Model-assisted survey regression estimators combine auxiliary information available at a population level with complex survey data to estimate finite population parameters. Many prediction methods, including linear and mixed models, nonparametric regression, and machine learning techniques, can be incorporated into such model-assisted estimators. These methods assume that observations obtained for the sample can be matched without error to the auxiliary data. We investigate properties of estimators that rely on matching algorithms that do not in general yield perfect matches. We focus on difference estimators, which are exactly unbiased under perfect matching but not under imperfect matching. The methods are investigated analytically and via simulation, using a study of recreational angling in South Carolina to build a simulation population. In this study, the survey data come from a stratified, two-stage sample and the auxiliary data from logbooks filed by boat captains. Extensions to multiple frame estimators under imperfect matching are discussed.

The second estimation problem involves the combination of survey data from a probability sample with additional data from a nonprobability sample. The problem is motivated by an application in which field crews are allowed to use their judgment in selecting part of a sample. Many surveys are conducted in two or more stages, with the first stage of primary sampling units dedicated to screening for secondary sampling units of interest, which are then measured or subsampled. The Large Pelagics Intercept Survey, conducted by the United States National Marine Fisheries Service, draws a probability sample of fishing access site-days in the first stage and screens for

relatively rare fishing trips that target pelagic species (tuna, sharks, billfish, etc.). Many site-days yield no pelagic trips. Motivated by this low yield, we consider surveys that allow expert judgment in the selection of some site-days. This nonprobability judgment sample is combined with a probability sample to generate likelihood-based estimates of inclusion probabilities and estimators of population totals that are related to dual-frame estimators. Consistency and asymptotic normality of the estimators are established under the correct specification of the model for judgment behavior. An extensive simulation study shows the robustness of the methodology to misspecification of the judgment behavior. A standard variance estimator, readily available in statistical software, yields stable estimates with small negative bias and good confidence interval coverage. Across a range of conditions, the proposed strategy that allows for some judgment dominates the classic strategy of pure probability sampling with known design weights. The methodology is extended to a doubly-robust version that uses both a propensity model for judgment selection probabilities and a regression model for study variable characteristics. If either model is correctly specified, the doubly-robust estimator is unbiased.

The dual-frame methodology for samples incorporating expert judgment is then extended to two other nonprobability settings: respondent-driven sampling and biased-frame sampling.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. F. Jay Breidt for his consistent support and guidance. I would also like to thank Dr. Haonan Wang, Dr. Joshua Keller and Dr. Sangmi Pallickara for their time and effort as committee members. Furthermore, I would like to thank NOAA Fisheries for their support and insightful discussion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Design-based estimation basics

Survey statistics consists of selecting a subset of a finite population and estimating something about the whole population based on that sample. Let us consider the finite population containing N elements, $U = \{1, 2, \ldots, N\}$, and let y_k denote the value of a nonrandom variable of interest for the kth element. We focus here on the estimation of the population total $T_y = \sum_{k \in U} y_k$. In the *design-based* approach to survey inference, all randomness comes from the random selection of the sample. Let $s \subset U$ denote a sample from the population selected via the sampling design $p(s)$, which is a probability distribution on the set of 2^N subsets of U if we include the empty set. The inclusion of the given element k in the sample is indicated by the random variable I_k, $I_k = 1$ if $k \in s$ and $I_k = 0$ if $k \notin s$. The probability that element k will be in the sample, denoted π_k, is computed from the design $p(\cdot)$ as $\pi_k = \mathrm{P}(k \in s) = \mathrm{E}[I_k] = \sum_{s \subset U : k \in s} p(s)$. The design is a *probability sampling design* if $\pi_k > 0$ for all $k \in U$. The probability that both of the elements k and ℓ will be included in the sample is $\pi_{k\ell} = \mathrm{P}(k, \ell \in s) = \mathrm{E}[I_k I_\ell] = \sum_{s \subset U : k, \ell \in s} p(s)$. In a probability sampling design, the Horvitz and Thompson (1952) estimator for the population total,

$$\widehat{T}_y = \sum_{k \in U} y_k \frac{I_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{\pi_k} \tag{1.1}$$

is design unbiased for T_y. The variance of the Horvitz-Thompson estimator is then

$$\mathrm{Var}\left[\widehat{T}_y\right] = \sum_{k, \ell \in U} \sum \mathrm{Cov}\left[I_k, I_\ell\right] \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} = \sum_{k, \ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}, \tag{1.2}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$. If $\pi_{k\ell} > 0$ for all $k, \ell \in U$, the design is a *measurable* design and an unbiased estimator of $\mathrm{Var}\left[\widehat{T}_y\right]$ is

$$\widehat{V}(\widehat{T}_y) = \sum_{k,\ell \in U} \sum \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}. \tag{1.3}$$

This unbiased estimator or (more commonly) approximations to it are computed in standard software such as the `survey` package of `R`.

1.2 Design-based model-assisted survey estimation

In many sampling situations, information about the study variable of the population may be available before sampling. The approach that incorporates additional information and model in the design-based estimation is called the design-based model-assisted estimation. We refer to the additional information as auxiliary variables and denote $\boldsymbol{x}_k$ for the vector of auxiliary variables for the element k. For the elements $k \in s$, $(y_k, \boldsymbol{x}_k)$ is observed and we assume the population total $\sum_{k \in U} \boldsymbol{x}_k$ is known. (That is, we do not need to know the disaggregated values $\boldsymbol{x}_k$ for $k \in U \setminus s$, but only for $k \in s$.)

Suppose the approximation of y_k using the auxiliary variable can be written as some method $m(\boldsymbol{x}_k)$ where the method $m(.)$ does not depend on the sample. The difference estimator for the population total is

$$\widehat{T}_{y,\text{diff}} = \sum_{k \in U} m(\boldsymbol{x}_k) + \sum_{k \in s} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k}.$$

Under a probability sampling design, the difference estimator is exactly unbiased for T_y regardless of the performance of $m(.)$. The variance is given by

$$\text{Var}\left[\widehat{T}_{y,\text{diff}}\right] = \sum_{k,\ell \in U} \sum \Delta_{k\ell} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell}.$$

An unbiased variance estimator under a measurable sampling design is

$$\widehat{V}\left(\widehat{T}_{y,\text{diff}}\right) = \sum_{k,\ell \in U} \sum \Delta_{k\ell} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}.$$

The difference estimator will have a smaller variance than Horvitz-Thompson if the residual $\{y_k - m(\boldsymbol{x}_k)\}$ has a smaller variation than $\{y_k\}$. When y_k is exactly $m(\boldsymbol{x}_k)$ for $k = 1, \ldots, N$, the difference estimator is completely error free.

In practice, it is rare to have a sample-independent $m(.)$ that is a good approximation of y_k. Instead, $m(\boldsymbol{x}_k)$ is often estimated based on the sample data $\{(y_k, \boldsymbol{x}_k)\}_{k \in s}$. Model-assisted estimators relax the condition that $m(.)$ is independent of the sample by introducing a model for which

$$\mathrm{E}\left[y_k\right] = \mu\left(\boldsymbol{x}_k\right),$$

where $\{y_k\}_{k \in U}$ are now assumed to be realized values from the superpopulation model. If we observe y_k and $\boldsymbol{x}_k$ for the entire population, we would get $m_N(.)$ as an estimate for $\mu(.)$. Since only the sample of y_k and $\boldsymbol{x}_k$ is observed, we estimate the model $\mu(.)$ from the sample and obtain $\widehat{m}(.)$. Plugging $\widehat{m}(.)$ into the difference estimator, we obtain the model-assisted estimator

$$\widehat{T}_{y,\mathrm{ma}} = \sum_{k \in U} \widehat{m}(\boldsymbol{x}_k) + \sum_{k \in s} \frac{y_k - \widehat{m}(\boldsymbol{x}_k)}{\pi_k}. \tag{1.4}$$

The model-assisted estimator is asymptotically unbiased regardless of the model quality. The variance is asymptotically equivalent to the corresponding difference estimator

$$\mathrm{Var}\left[\widehat{T}_{y,\mathrm{diff}}\right] = \sum_{k,\ell \in U} \sum \Delta_{k\ell} \frac{y_k - m_N(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m_N(\boldsymbol{x}_\ell)}{\pi_\ell}.$$

The variance estimator is

$$\widehat{V}\left(\widehat{T}_{y,\mathrm{ma}}\right) = \sum_{k,\ell \in U} \sum \Delta_{k\ell} \frac{y_k - \widehat{m}(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - \widehat{m}(\boldsymbol{x}_\ell)}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}.$$

Similarly to the difference estimator, the asymptotic variance is smaller than the Horvitz-Thompson estimator if the residual $\{y_k - m_N(\boldsymbol{x}_k)\}$ has less variation than $\{y_k\}$.

An important case of the model-assisted estimator is the generalized linear regression estimators (GREG) (Särndal et al., 1992) with the working model

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta} + \varepsilon_k,$$

where ε_k is uncorrelated with mean 0 and variance σ_k^2. All σ_k^2 are known. If we observe y_k and $\boldsymbol{x}_k$ for all k in U, the weighted least square estimator of $\boldsymbol{\beta}$ under the model is

$$\boldsymbol{B}_N = \left(\sum_{k \in U} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{\boldsymbol{x}_k y_k}{\sigma_k^2}.$$

But we can estimate (1.2) using the sample data by plugging in HT estimators,

$$\widehat{\boldsymbol{B}} = \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k \in s} \frac{\boldsymbol{x}_k y_k}{\sigma_k^2 \pi_k}.$$

The corresponding predictor would be

$$\widehat{m}\left(\boldsymbol{x}_k\right) = \boldsymbol{x}_k^\top \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k \in s} \frac{\boldsymbol{x}_k y_k}{\sigma_k^2 \pi_k}. \tag{1.5}$$

By plugging (1.5) into the model-assisted estimator from (1.4), we have

$$\widehat{T}_{y,\mathrm{GREG}} = \sum_{k \in U} \boldsymbol{x}_k^\top \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k \in s} \frac{\boldsymbol{x}_k y_k}{\sigma_k^2 \pi_k} + \sum_{k \in s} \left\{ y_k - \boldsymbol{x}_k^\top \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k \in s} \frac{\boldsymbol{x}_k y_k}{\sigma_k^2 \pi_k} \right\} \left(\pi_k\right)^{-1}.$$

GREG estimator is also asymptotically unbiased regardless of the working model, and the asymptotic variance is

$$\mathrm{Var}\left[\widehat{T}_{y,\mathrm{GREG}}\right] = \sum_{k,\ell \in U} \sum \Delta_{k\ell} \frac{y_k - \boldsymbol{x}_k^\top \boldsymbol{B}_N}{\pi_k} \frac{y_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{B}_N}{\pi_\ell}.$$

1.3 Dual-frame estimation

In sampling theory, we often assume the frame is complete. However, a complete and perfect frame is often not feasible in practice. Multiple-frame techniques use two or more frames to give complete coverage of the target population and thus deal with the imperfect frames. Suppose every element k is in at least one of the two frames U_A and U_B, and two imperfect frames provide the complete coverage of the target population $U = U_A \cup U_B$. A probability sample is taken from each frame independently, s_A by a probability sampling design p_A from U_A, and s_B by a probability sampling design p_B from U_B. Let N_A be the size of U_A, and N_B is the size of U_B. The Horvitz-Thompson estimators in each frame are $\widehat{T}_A = \sum_{k \in s_A} y_k/\pi_k^A$ and $\widehat{T}_B = \sum_{k \in s_B} y_k/\pi_k^B$. The estimators are design unbiased for their respective frame totals: $\mathrm{E}\left[\widehat{T}_A\right] = T_A = \sum_{k \in U_A} y_k$, and $\mathrm{E}\left[\widehat{T}_B\right] = T_B = \sum_{k \in U_B} y_k$. We also distinguish three disjoint domains of the population U: $a = U \setminus U_B$, $b = U \setminus U_A$, and $ab = U_A \cap U_B$ with sizes N_a, N_b, and N_{ab}. Domain a consists of the elements in U_A only, domain b consists of the elements in U_B only, and domain ab consists of the elements in both U_A and U_B. We give an overview of the combined frame approach, separate frame approach, and multiplicity estimators. The approaches we describe generalize to more than two frames.

1.3.1 Combined frame estimators

The combined approach combines samples from all frames into a single combined sample with appropriate weights. It is closely related to the estimators in Bankier (1986) and Kalton and Anderson (1986). In the multiple frame literature, this approach was sometimes referred to as a single frame estimator (Lohr, 2009). We will refer to this approach as the combined frame estimator to avoid ambiguity with the traditional single frame estimator. The inclusion probability of the combined sample is

$$\mathrm{P}\left[k \in s\right] = \mathrm{P}\left[k \in s_A \cup s_B\right] = \mathrm{P}\left[k \in s_A\right] + \mathrm{P}\left[k \in s_B\right] - \mathrm{P}\left[k \in s_A \cap s_B\right] \simeq \pi_k^A + \pi_k^B.$$

The approximation improves as the sampling fractions in both frames decrease. The combined frame estimator for the population total is

$$\widehat{T}_{\text{com}} = \sum_{k \in s} \frac{y_k}{\pi_k^A + \pi_k^B},$$

assuming no element occurs twice in the combined sample. The estimator is approximately unbiased if the sampling fraction is small. Because the combined estimator can be written as $\sum_{k \in s} y_k w_k$ and the two designs are independent, the variance and variance estimator can be derived from the usual design-based variance estimation. However, there are some drawbacks of the combined frame estimator. The estimator requires the inclusion probability for every sampled unit for all frames, not just the inclusion probability for the frame from which the element was selected.

1.3.2 Separate frame estimators

The separate approach first computes estimates for each of the disjoint domains and then aggregates all over all domains. It is the largest class of estimators for multiple frame surveys. In the setting of the two imperfect frames, we compute the estimators

$$\widehat{T}_a^A = \sum_{k \in s_A} \frac{y_k \mathbf{1}_{\{k \in a\}}}{\pi_k^A}; \widehat{T}_{ab}^A = \sum_{k \in s_A} \frac{y_k \mathbf{1}_{\{k \in ab\}}}{\pi_k^A}; \widehat{T}_{ab}^B = \sum_{k \in s_B} \frac{y_k \mathbf{1}_{\{k \in ab\}}}{\pi_k^B}; \widehat{T}_b^B = \sum_{k \in s_B} \frac{y_k \mathbf{1}_{\{k \in b\}}}{\pi_k^B}$$

and the estimation for the population total is

$$\widehat{T}_{\text{sep}} = \widehat{T}_a^A + \alpha \widehat{T}_{ab}^A + (1 - \alpha) \widehat{T}_{ab}^B + \widehat{T}_b^B \tag{1.6}$$

with $\alpha \in [0, 1]$. The estimator only requires knowledge about the frame membership for the sampled unit and the inclusion probabilities for the sampled unit for the frames from which they were selected. When $\alpha = 0$ or $\alpha = 1$, the estimator becomes the screening estimator in which the elements in the overlap domain ab from one of the samples are screened out. The optimal value of

α is chosen to minimize the variance of $\widehat{T}_{\text{sep}}$ and is given by

$$\alpha_{\text{opt}} = \frac{\text{Var}\left(\widehat{T}_{ab}^{B}\right) + \text{Cov}\left(\widehat{T}_{b}^{B}, \widehat{T}_{ab}^{B}\right) - \text{Cov}\left(\widehat{T}_{a}^{A}, \widehat{T}_{ab}^{A}\right)}{\text{Var}\left(\widehat{T}_{ab}^{A}\right) + \text{Var}\left(\widehat{T}_{ab}^{B}\right)}.$$

The variance and covariance are unknown and have to be estimated from the sample. The optimal value for one survey variable might be different from the optimal value for another survey variable. Alternatives to the optimal estimator include Fuller and Burmeister (1972), who proposed a regression-type estimator by using information about N_{ab}, and Skinner and Rao (1996), who proposed a pseudo maximum likelihood (PML) estimator that has the same weights for all the variables. These modified estimators are not considered further in this dissertation.

1.3.3 Multiplicity estimators

The idea of multiplicity was introduced in Casady and Sirken (1980) and described in Mecatti (2007). Instead of creating the single design, the multiplicity estimator counts the number of frames that include unit k: its multiplicity, m_k. The estimator for two frames is

$$\widehat{T}_{y,\text{mec}} = \sum_{k \in s_A} \frac{y_k}{m_k \pi_k^A} + \sum_{k \in s_B} \frac{y_k}{m_k \pi_k^B}. \tag{1.7}$$

In the two-frame case, the multiplicity $m_k = 1$ for $k \in a$ and $k \in b$, and $m_k = 2$ for $k \in ab$, so that (1.7) is a special case of the separate frame estimator (1.6) with $\alpha = 1/2$.

The multiplicity estimator does not involve the domain membership indicator that identifies which domain includes element k, hence it is simple to implement. Like the separate frame estimator, it only requires knowledge of the inclusion probabilities of the sample selected. The variance and variance estimator can be derived in closed form:

$$\text{Var}\left[\widehat{T}_{y,\text{mec}}\right] = \sum_{k,\ell \in U_A} \sum \Delta_{k\ell} \frac{y_k}{m_k \pi_k^A} \frac{y_\ell}{m_\ell \pi_\ell^A} + \sum_{k,\ell \in U_B} \sum \Delta_{k\ell} \frac{y_k}{m_k \pi_k^B} \frac{y_\ell}{m_\ell \pi_\ell^B},$$

$$\widehat{V}(\widehat{T}_{y,\text{mec}}) = \sum_{k,\ell \in U_A} \sum \Delta_{k\ell} \frac{y_k}{m_k \pi_k^A} \frac{y_\ell}{m_\ell \pi_\ell^A} \frac{I_k^A I_\ell^A}{\pi_{k\ell}^A} + \sum_{k,\ell \in U_B} \sum \Delta_{k\ell} \frac{y_k}{m_k \pi_k^B} \frac{y_\ell}{m_\ell \pi_\ell^B} \frac{I_k^B I_\ell^B}{\pi_{k\ell}^B}. \tag{1.8}$$

1.4 Inference for nonprobability samples

With the increasing cost and lower response rate of the traditional probability sample surveys, the low cost of nonprobability samples of very large size through web surveys and external sources like administrative data has been really attractive. Estimation involves nonprobability samples that may be biased due to unknown selection mechanism. There are some approaches in the literature that discuss inference for the nonprobability samples that combine probability and nonprobability samples. We give an overview of four main classes of estimation: small area estimation, sample matching and mass imputation, inverse weighted estimation, and doubly-robust estimation. To apply uniform notation for these classes, let s_A denote the probability sample with sample size n_A, and s_B denote the nonprobability sample with sample size n_B. The probability sample indicators are I_k^A if $k \in s_A$, $I_k^A = 0$ otherwise; similarly, the nonprobability sample indicators are I_k^B if $k \in s_B$, $I_k^B = 0$ otherwise. The auxiliary variable $\boldsymbol{x}_k$ is observed (at least) in both probability and nonprobability samples, and the variable of interest y_k is observed in at least one of the two samples.

1.4.1 Small area estimation

The small area approach uses the bivariate Fay-Herriot models to get the domain level point estimate of probability samples and nonprobability samples (Ganesh et al., 2017). The models for the domain are

$$\begin{aligned} y_d^A &= \alpha_d + \nu_d + e_d^A \\ y_d^B &= \alpha_d + \beta_d + \nu_d + e_d^B, \end{aligned}$$

where y_d^A, y_d^B are the point estimates from the probability sample and nonprobability sample for domain d; α_d, β_d are domain level fixed effects; ν_d are domain level random effects with $\nu_d \sim$

$\mathrm{N}(0, \sigma_\nu^2)$; and e_d^A and e_d^B are sampling errors associated with y_d^A, y_d^B, with $e_d^A \sim \mathrm{N}(0, \psi_d^A)$ and $e_d^B \sim \mathrm{N}(0, \psi_d^B)$. The unknown parameters in the model are estimated by the maximum likelihood estimator. The drawback of this estimator is it depends on the variable being estimated and needs a different model for different variables of interest.

1.4.2 Sample matching and mass imputation

The mass imputation approach is a model-based prediction method (Chen et al., 2020; Kim et al., 2021). Suppose the finite population can be viewed as a sample from the superpopulation model

$$y_k = m(\boldsymbol{x}_k) + \varepsilon_k, k = 1, \ldots, N,$$

where $m(\boldsymbol{x}_k) = \mathrm{E}[y_k | \boldsymbol{x}_k]$. The method fits the model and estimates the coefficients from the nonprobability sample s_B, and the mass imputation estimator of the population total is obtained by the weighted sum of the predicted values $\widehat{m}(\boldsymbol{x}_k)$ over the probability sample s_A:

$$\sum_{k \in s_A} \frac{\widehat{y}_k}{\pi_k^A}. \tag{1.9}$$

The estimator can be viewed as replacing the missing response variable in the probability sample by the imputed value from the observed covariates, so that we obtain predicted values $\widehat{y}_k$ for $k \in s_A$. The mass imputation estimator (1.9) performs poorly if the superpopulation model is not correctly specified.

The sample matching approach uses a non-parametric approach without specifying the parametric model $m(\boldsymbol{x}_k)$. The sample matching approach assigns sampling weights from the probability sample to the nonprobability sample by comparing covariates available in both samples using distance measure with the nearest neighbor (Chen et al., 2020; Rivers, 2007; Yang et al., 2018). This approach resembles the donor from the probability sample to a recipient from the nonprobability sample, and the weight of the nonprobability sample element is from the nearest probability sample element.

1.4.3 Inverse weighted estimator

The inverse weighting approach fits a model for the inclusion probabilities of elements in the nonprobability sample s_B based on the missing at random assumption. In the nonprobability literature, this approach was discussed by several authors, Kim and Wang (2019), Chen et al. (2020), Elliott and Valliant (2017), and Valliant (2020). Suppose the selection mechanism for the nonprobability sample, referred to as propensity score, is $\pi_k^B = \mathrm{P}\left[I_k^B = 1 \middle| \boldsymbol{x}_k, y_k\right], k \in U$. Further assume that the selection mechanism is ignorable in the sense that $\pi_k^B = \mathrm{P}\left[I_k^B = 1 \middle| \boldsymbol{x}_k, y_k\right] = \mathrm{P}\left[I_k^B = 1 \middle| \boldsymbol{x}_k\right]$ for all k. Suppose the mechanism follows a parametric model $\mathrm{P}\left[I_k^B = 1 \middle| \boldsymbol{x}_k\right] = p_k\left(\boldsymbol{\theta}\right)$, so that the log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{k \in U}\left[I_k^B \log p_k + (1 - I_k^B)\log(1 - p_k)\right] = \sum_{k \in U} I_k^B \log\left[\frac{p_k(\boldsymbol{\theta})}{1 - p_k(\boldsymbol{\theta})}\right] + \sum_{k \in U}\log\left(1 - p_k\left(\boldsymbol{\theta}\right)\right),$$

which involves the unknown and unobservable term $\log\left(1 - p_k\left(\boldsymbol{\theta}\right)\right)$. This unknown term is replaced by its design unbiased estimator based on the probability sample s_A. The procedure leads to the pseudo log-likelihood

$$l^*(\boldsymbol{\theta}) = \sum_{k \in s_B}\log\left(\frac{p_k\left(\boldsymbol{\theta}\right)}{1 - p_k\left(\boldsymbol{\theta}\right)}\right) + \sum_{k \in s_A}\frac{\log\left(1 - p_k\left(\boldsymbol{\theta}\right)\right)}{\pi_k^A},$$

and $\widehat{\boldsymbol{\theta}}$ is computed by maximizing the pseudo log-likelihood function $l^*(\boldsymbol{\theta})$. The estimated propensity score $\widehat{\pi}_k$ is then obtained from the fitted model and the estimator for the population total is $\widehat{T}_y = \sum_{k \in s_B} y_k / \widehat{\pi}_k^B$. The variance of the estimator can be obtained by the joint estimating equation of estimating the total and the score function of the pseudo log-likelihood (Kim and Wang (2019), Chen et al. (2020)). The drawback of the inverse weighted estimator is that this estimator is sensitive to the misspecified models for the propensity scores.

1.4.4 Doubly-robust estimator

Due to the possible misspecified models for the propensity scores, the estimators can be improved by including a predictive model. This approach introduces a regression model $\mathrm{E}\left[y \middle| \boldsymbol{x}\right] =$

$m(\boldsymbol{x}, \boldsymbol{\beta})$ together with the propensity score to construct the estimator. The general form of the doubly-robust estimator for the total is

$$\widehat{T}_{y,\mathrm{DR}} = \sum_{k \in s_B} \frac{y_k - \widehat{m}(\boldsymbol{x}_k)}{\widehat{\pi}_k^B} + \sum_{k \in U} \widehat{m}(\boldsymbol{x}_k). \tag{1.10}$$

The coefficients of $\widehat{m}(\boldsymbol{x}_k)$ can be obtained by the standard methods such as least squares or maximum likelihood from the nonprobability sample only. The estimator in (1.10) is similar to the model-assisted estimator. The doubly-robust estimator would remain consistent if either the model for propensity score or the predictive model is correctly specified. Doubly-robust estimator have been used in the context of missing data (Kim and Haziza (2014), Carpenter et al. (2006), Davidian et al. (2005), Bang and Robins (2005), Kang and Schafer (2007)).

1.5 Respondent-driven sampling (RDS)

Respondent-driven sampling (RDS) was proposed to sample and produce unbiased estimates for hidden or rare populations. Some studies of such populations use link-tracing designs, in which links or connections between units are used in obtaining the sample (Thompson, 2012), because these designs can exploit relationships among the rare units. But such designs are subject to bias because the search procedure or the design is not taken into account. The RDS approach allows for statistical inference for the population by controlling the bias that is often associated with the link-tracing design. Each individual is given a limited number of coupons to recruit acquaintances who are then interviewed. Each new respondent is in turn given coupons to recruit acquaintances. The process continues until reaching the target sample size. From each respondent, data are collected on variables of interest and the number of connections of the respondent, or *degree* of that individual. There is an extensive literature on inference for respondent-driven sampling, including the overview such as Heckathorn and Cameron (2017). We give an overview of the most widely used classes of RDS estimators: SH (Salganik and Heckathorn (2004)), VH (Volz and Heckathorn (2008)), and SS (successive sampling; Gile (2011)).

1.5.1 Salganik and Heckathorn (SH) estimator

Salganik and Heckathorn (2004) derived the SH estimator by Markov modeling and the information about the mean degree (personal network size) of each group. Assuming that seeds are drawn with probability proportional to degree, seeds recruit randomly within their networks, and under additional mild assumptions, the probability that a node j will be recruited is $d_j / \sum_{k \in N} d_k$, where d_j is the network degree of person j, that is, the nodes in the next waves will be drawn proportional to degree, and each relationship has the same probability of being drawn with $1/\sum_{k \in N} d_k$. The estimator estimates the proportion of the population in two groups, A and B,

$$\widehat{\mu}_A^{\text{SH}} = \frac{\widehat{d}_B \widehat{C}_{BA}}{\widehat{d}_A \widehat{C}_{AB} + \widehat{d}_B \widehat{C}_{BA}},$$

where $\widehat{d}_A$, $\widehat{d}_B$ are the estimated average degrees in group A and group B, and $\widehat{C}_{AB}$, $\widehat{C}_{AB}$ are the estimated probabilities of cross-group recruitment. The estimated average degree in group A is

$$\widehat{d}_A = \frac{1/n_A \sum_{k=1}^{n_A} d_k 1/p_k}{1/n_A \sum_{k=1}^{n_A} 1/p_k},$$

where p_k is the selection probability of person k and n_A is the sample size of group A. Because nodes are drawn with probability proportional to degree, the estimated degree reduces to

$$\widehat{d}_A = \frac{n_A}{\sum_{k=1}^{n_A} 1/d_k}.$$

Cross-group recruitment probability is estimated by

$$\widehat{C}_{AB} = \frac{r_{AB}}{r_{AA} + r_{AB}}; \widehat{C}_{BA} = \frac{r_{BA}}{r_{BB} + r_{BA}},$$

where r_{AA} is the number of recruitments from a person in group A to another person in group A, r_{AB} is the number of recruitments from a person in group A to another person in group B, r_{BA} is

the number of recruitments from a person in group B to another person in group A, and r_{BB} is the number of recruitments from a person in group B to another person in group B.

1.5.2 Volz and Heckathorn (VH) estimator

The estimator was proposed by Volz and Heckathorn (2008) and is most commonly used today. This estimator is similar to SH estimator but has a different theoretical foundation. Recruitment is modeled as a Markov process (MP) and assumes the existence of a unique equilibrium to the MP. The VH estimator calculates the selection probability and utilize the Hansen-Hurwitz (HH) type estimator (Hansen and Hurwitz, 1943). From the equilibrium of MP, the selection probability is estimated as $p_k = d_k/N\widehat{d}_U$, where $\widehat{d}_U$ is the estimate of the average degree. From Salganik and Heckathorn (2004), the estimated $\widehat{d}_U$ is $n/\left(\sum_{k\in s} 1/d_k\right)$. The estimate for the total is then

$$\widehat{T}_y^{\mathrm{VH}} = \frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{\widehat{p}_k} = \frac{1}{n}\sum_{k=1}^{n}\frac{Ny_k n/\sum_{i\in s} d_k^{-1}}{d_k^{-1}}.$$

The estimate for the mean is

$$\widehat{\mu}_y^{\mathrm{VH}} = \frac{\sum_{k\in s} d_k^{-1} y_k}{\sum_{k\in s} d_k^{-1}},$$

where d_k is the degree of element k. SH estimator and VH estimator will coincide if the number of recruitments from Group A to Group B is the same as recruitments from Group B to Group A. VH estimator also requires some mild assumptions as in the SH estimator but allows estimation for the continuous variable as opposed to SH estimator, which only allows for estimation for categorical variables. Gile and Handcock (2010) conducted the simulation of VH and SH estimators under realistic sampling scenarios which violated the assumptions like random seeds or sampling with replacement, and showed that such violations can affect the estimates.

1.5.3 Successive sampling (SS) estimator

Gile (2011) proposed the successive sampling (SS) estimator that avoids the bias from with-replacement assumption and the assumption of reaching the stationary equilibrium of the VH es-

timator. The estimator is based on the sequential sample, and the idea about the successive sampling is given the population degree distribution and the sample size, there is unknown function $f_\pi\left(k; n, \mathcal{N}\right)$ mapping the degree of the individual to its inclusion probability π_k under without replacement sampling. The mapping function depends on the population size N, which is assumed known, and the degree distribution $\mathcal{N}$. This approach estimates the degree distribution $\mathcal{N}$ and the mapping function iteratively. The resulting estimator for the population mean is

$$\widehat{\mu}_y^{\text{SS}} = \frac{\sum_{k\in s} y_k/\widehat{\pi}(d_k)}{\sum_{k\in s} 1/\widehat{\pi}(d_k)},$$

where $\widehat{\pi}(d_k)$ is the estimated inclusion probability with degree d_k. The estimator fixes the bias for the large sampling fraction in the VH estimator and it performs similarly to the VH estimator if the sampling fraction is small.

1.6 Overview of organization of thesis

We address two main topics in the dissertation, imperfect matching and combining probability and nonprobability samples, with some extensions. In Chapter 2, we introduce the imperfectly matched model-assisted estimator, in which the sample is matched with error to the auxiliary information. The methods are investigated analytically and via simulation, using a study of recreational angling in South Carolina to build a simulation population. The first part of Chapter 2, from section 2.1 to 2.3 and part of 2.5, has been published in the refereed volume in Breidt et al. (2018). Chapter 3 is the topic about the inference incorporating expert judgment sample. Motivated by the low yield of the Large Pelagics Intercept Survey, a two-stage screening sample for a rare type of recreational fishing activity, we have considered surveys that allow expert judgment in the selection of some primary sampling units. This nonprobability judgment sample is combined with a probability sample to generate likelihood-based estimates of inclusion probabilities and estimators of population totals that are related to dual-frame estimators. Consistency and asymptotic normality of the estimators are established under the correct specification of the model for judgment behavior.

Chapter 4 is the extension and application of Chapter 3 to other problems including doubly-robust estimation, estimation for RDS samples, and incomplete frame inference. A brief summary and discussion are given in Chapter 5.

Chapter 2

Model-assisted survey estimation with imperfect matching

2.1 Introduction

Let $U = \{1, 2, \ldots, N\}$ denote a finite population and let y_k denote the (non-random) value of some variable of interest for element $k \in U$. We are interested in the finite population total $T_y = \sum_{k \in U} y_k$. As a motivating example, which we return to in section 2.5, suppose U is the set of all recreational angling boat trips on the coast of the state of South Carolina in 2016. Further, suppose y_k is the number of anglers on the kth boat trip, so that T_y is the total number of recreational angler trips on boats in South Carolina waters in 2016; or suppose that y_k is the number of black sea bass caught on the kth boat trip, so that T_y is the total number of black sea bass caught in 2016. Because it is often impractical to measure y_k for all $k \in U$, we instead estimate T_y based on information obtained for a sample $s \subset U$, which is selected via a random mechanism.

In addition to observations obtained on the sample, auxiliary information may be available from external records. Let $\mathcal{A} = \{1, 2, \ldots, A\}$ denote the indices for this external database and let $\boldsymbol{a}_\ell$ denote the vector of auxiliary information available for record $\ell \in \mathcal{A}$. We write $A_x = \sum_{\ell \in \mathcal{A}} x_\ell$ for sums over the database, in particular noting that the size of the database is $A_1 = \sum_{\ell \in \mathcal{A}} 1$.

The auxiliary vector $\boldsymbol{a}_\ell$ could be used to construct a predictor, $\mu(\boldsymbol{a}_\ell)$ of y_k provided record $\ell \in \mathcal{A}$ in the database matches element $k \in U$ in the population. We assume for the present that the construction of the prediction method $\mu(\cdot)$ does not involve the sample, s. We write $A_\mu = \sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell)$. In the motivating example, we have samples from angler interviews and logbook records as auxiliary information.

2.2 Estimation under perfect matching

2.2.1 Notation for perfect matching

We first consider the case of perfect matching: suppose that every record in the database can be matched to one and only one element in the population, and vice versa. We write

$$M_{k\ell} = \begin{cases} 1, & \text{if database record } \ell \in \mathcal{A} \text{ matches element } k \in U, \\ 0, & \text{otherwise.} \end{cases}$$

The appropriate predictor of y_k would then be denoted $\sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{a}_\ell)$, to reflect the matching step.

2.2.2 Difference estimator under perfect matching.

Under this perfect matching scenario, $\sum_{k \in U} M_{k\ell} = 1$. It follows that an unbiased estimator of T_y is given by the *difference estimator*,

$$\begin{aligned} \widehat{T}_{y,\text{diff}} &= \sum_{k \in U}\sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{a}_\ell) + \sum_{k \in s} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{a}_\ell)}{\pi_k} \\ &= \sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell) + \sum_{k \in s} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{a}_\ell)}{\pi_k}; \end{aligned} \tag{2.1}$$

this is simply a more elaborate notation for a standard estimator (e.g., equation (4) of Breidt and Opsomer (2017)), to account for the matching step. The variance of the perfect-matching difference estimator is

$$\text{Var}\left[\widehat{T}_{y,\text{diff}}\right] = \sum_{j,k \in U}\sum \Delta_{jk} \frac{y_j - \sum_{\ell \in \mathcal{A}} M_{j\ell}\mu(\boldsymbol{a}_\ell)}{\pi_j} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{a}_\ell)}{\pi_k}. \tag{2.2}$$

The unbiased difference estimator will have smaller variance and mean square error than the unbiased Horvitz-Thompson estimator (1.1) provided the residuals $\{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell}\mu(\boldsymbol{a}_\ell)\}_{k \in U}$ in (2.2) have less variation than the raw values $\{y_k\}_{k \in U}$ in (1.2).

2.3 Estimation under imperfect matching

2.3.1 Notation for imperfect matching

In practice, perfect matching may not be possible. The sampled element k might have no corresponding record in the database. It might have a corresponding record ℓ, but fail to match it perfectly due to missing values or inaccuracies in the survey observation, the database record, or both. Similarly, the sampled element might appear to match multiple database records due to agreement on a number of data values.

Hence, we replace the $M_{k\ell} = 0$ or 1 by a possibly-fractional value $m_{k\ell} \in [0, 1]$, computed via a deterministic algorithm that does not depend on the sample. We refer to these values as *match metrics*. Assume that for any sampled element $k \in U$, the match metrics $\{m_{k\ell}\}_{\ell \in \mathcal{A}}$ for every database record can be computed. For example, sampled element k might match record ℓ_1 perfectly, in which case

$$m_{k\ell} = \begin{cases} 1, & \text{if } \ell = \ell_1, \\ 0, & \text{otherwise.} \end{cases}$$

It might not match any records, in which case

$$m_{k\ell} = 0, \quad \text{for all } \ell \in \mathcal{A};$$

or it might match three records ℓ_1, ℓ_2, ℓ_3 equally well, in which case

$$m_{k\ell} = \begin{cases} 1/3, & \text{if } \ell = \ell_1, \ell = \ell_2 \text{ or } \ell = \ell_3, \\ 0, & \text{otherwise.} \end{cases}$$

If $\sum_{\ell \in \mathcal{A}} m_{k\ell} < 1$, then the matching algorithm has determined that there is a non-trivial possibility that the sampled element does not match any database record. This can occur when there is potential non-overlap between the target population U and the database $\mathcal{A}$. This is of interest in situations such as the application we will describe in Section 2.5, where $\mathcal{A}$ is a possibly-incomplete

set of recreational angling trips self-reported by boat captains, while U is the actual population of trips.

2.3.2 Difference estimation under imperfect matching

First difference estimator

Under imperfect matching, an estimator analogous to (2.1) is

$$\widetilde{T}_{y,\text{diff1}} = \sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell) + \sum_{k \in s} \frac{y_k - \sum_{\ell \in \mathcal{A}} m_{k\ell} \mu(\boldsymbol{a}_\ell)}{\pi_k}. \tag{2.3}$$

This estimator is no longer unbiased. Instead, its expectation is

$$\begin{aligned} \mathrm{E}\left[\widetilde{T}_{y,\text{diff1}}\right] &= \sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell) + \sum_{k \in U} \left(y_k - \sum_{\ell \in \mathcal{A}} m_{k\ell} \mu(\boldsymbol{a}_\ell) \right) \\ &= T_y + \sum_{\ell \in \mathcal{A}} \left(1 - \sum_{k \in U} m_{k\ell} \right) \mu(\boldsymbol{a}_\ell). \end{aligned} \tag{2.4}$$

Its variance is

$$\mathrm{Var}\left[\widetilde{T}_{y,\text{diff1}}\right] = \sum_{j,k \in U} \sum \Delta_{jk} \frac{y_j - \sum_{\ell \in \mathcal{A}} m_{j\ell} \mu(\boldsymbol{a}_\ell)}{\pi_j} \frac{y_k - \sum_{\ell \in \mathcal{A}} m_{k\ell} \mu(\boldsymbol{a}_\ell)}{\pi_k}.$$

The variance is small if the match-weighted quantities $\{\sum_{\ell \in \mathcal{A}} m_{k\ell} \mu(\boldsymbol{a}_\ell)\}_{k \in U}$ are good predictors of the response values $\{y_k\}_{k \in U}$. Under a measurable sampling design, an unbiased variance estimator is given by

$$\widehat{V}(\widetilde{T}_{y,\text{diff1}}) = \sum_{j,k \in U} \sum \Delta_{jk} \frac{y_j - \sum_{\ell \in \mathcal{A}} m_{j\ell} \mu(\boldsymbol{a}_\ell)}{\pi_j} \frac{y_k - \sum_{\ell \in \mathcal{A}} m_{k\ell} \mu(\boldsymbol{a}_\ell)}{\pi_k} \frac{I_j I_k}{\pi_{jk}} \tag{2.5}$$

which, like (1.3), can be computed or closely approximated using standard survey software.

The behavior of the estimator under three extreme cases is of interest. First, if there is no matching at all, so that $m_{k\ell} \equiv 0$ for all $k \in U$, $\ell \in \mathcal{A}$, then $\widetilde{T}_{y,\text{diff1}}$ becomes

$$\sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell) + \sum_{k \in s} \frac{y_k}{\pi_k} = A_\mu + \widehat{T}_y,$$

with expectation $A_\mu + T_y$ and variance equal to that of the Horvitz-Thompson estimator, (1.2). Effectively, the estimator regards the sampling design as having failed to cover the complete population, which is actually the disjoint union $\mathcal{A} \cup U$ and not U. It thus separately estimates the totals for the database and the universe and adds them together.

The second extreme case is that of full matching in the sense that $\sum_{k \in U} m_{k\ell} = 1$ for all $\ell \in \mathcal{A}$ (this is not the same as perfect matching). In this case, $\widetilde{T}_{y,\text{diff1}}$ is exactly unbiased for T_y by (2.4).

The third and final extreme case can occur if a rare characteristic appears in the population but is never encountered in the sample, so that $y_k \equiv 0$ for all $k \in s$. In this case, the estimator (2.3) becomes

$$\sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell) - \sum_{k \in s} \sum_{\ell \in \mathcal{A}} \frac{m_{k\ell} \mu(\boldsymbol{a}_\ell)}{\pi_k}. \tag{2.6}$$

This behavior may be undesirable, as for a non-negative characteristic with non-negative predictions, the estimator predicts less than what is known to be present in the database. This behavior is better than that of the Horvitz-Thompson estimator, however, which would estimate zero for the population with such a degenerate sample. Nonetheless, other difference-type estimators are worth considering, including the one proposed below.

Second difference estimator

An alternative to $\widetilde{T}_{y,\text{diff1}}$ in (2.3) is obtained by an additional differencing adjustment,

$$\begin{aligned} \widetilde{T}_{y,\text{diff2}} &= \widetilde{T}_{y,\text{diff1}} + \sum_{k \in s} \sum_{\ell \in \mathcal{A}} \frac{m_{k\ell} \left\{ \mu(\boldsymbol{a}_\ell) - y_k \right\}}{\pi_k} \\ &= \sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell) + \sum_{k \in s} \frac{y_k (1 - \sum_{\ell \in \mathcal{A}} m_{k\ell})}{\pi_k}. \end{aligned} \tag{2.7}$$

The expectation of the estimator is

$$\begin{aligned} \mathrm{E}\left[\widetilde{T}_{y,\text{diff2}}\right] &= \mathrm{E}\left[\widetilde{T}_{y,\text{diff1}}\right] + \sum_{k\in U}\sum_{\ell\in\mathcal{A}} m_{k\ell}\left\{\mu(\boldsymbol{a}_\ell) - y_k\right\} \\ &= T_y + \sum_{\ell\in\mathcal{A}}\left(1 - \sum_{k\in U} m_{k\ell}\right)\mu(\boldsymbol{a}_\ell) + \sum_{k\in U}\sum_{\ell\in\mathcal{A}} m_{k\ell}\left\{\mu(\boldsymbol{a}_\ell) - y_k\right\}. \end{aligned} \tag{2.8}$$

Its variance is

$$\mathrm{Var}\left[\widetilde{T}_{y,\text{diff2}}\right] = \sum_{j,k\in U}\sum \Delta_{jk}\frac{y_j(1-\sum_{\ell\in\mathcal{A}} m_{j\ell})}{\pi_j}\frac{y_k(1-\sum_{\ell\in\mathcal{A}} m_{k\ell})}{\pi_k}.$$

The variance is small if the matching is good in the sense that $\sum_{\ell\in\mathcal{A}} m_{k\ell} \simeq 1$ for all $k \in U$. Under a measurable sampling design, an unbiased variance estimator is given by

$$\widehat{V}(\widetilde{T}_{y,\text{diff2}}) = \sum_{j,k\in U}\sum \Delta_{jk}\frac{y_j(1-\sum_{\ell\in\mathcal{A}} m_{j\ell})}{\pi_j}\frac{y_k(1-\sum_{\ell\in\mathcal{A}} m_{k\ell})}{\pi_k}\frac{I_j I_k}{\pi_{jk}}. \tag{2.9}$$

Again, like (1.3) and (2.5), this estimator can be computed or closely approximated using standard survey software.

We next consider the behavior of $\widetilde{T}_{y,\text{diff2}}$ under the three extreme scenarios described above. First, if there is no matching at all, so that $m_{k\ell} \equiv 0$ for all $k \in U$, $\ell \in \mathcal{A}$, then $\widetilde{T}_{y,\text{diff2}}$ reduces to $\widetilde{T}_{y,\text{diff1}}$ by (2.7) and has exactly the same behavior.

Second, under full database matching in the sense that $\sum_{k\in U} m_{k\ell} = 1$ for all $\ell \in \mathcal{A}$, the expectation of $\widetilde{T}_{y,\text{diff2}}$ in (2.8) becomes

$$T_y + \sum_{\ell\in\mathcal{A}}\mu(\boldsymbol{a}_\ell) - \sum_{k\in U}\sum_{\ell\in\mathcal{A}} m_{k\ell}y_k$$

so that, unlike $\widetilde{T}_{y,\text{diff1}}$ under this scenario, $\widetilde{T}_{y,\text{diff2}}$ is biased. The bias is small if $\sum_{k\in U} m_{k\ell}y_k$ is close to $\mu(\boldsymbol{a}_\ell)$ for all ℓ.

Third, with $y_k \equiv 0$ for all $k \in s$, the estimate computed from (2.7) becomes the full database total

$$\sum_{\ell \in \mathcal{A}} \mu(\boldsymbol{a}_\ell),$$

which may be preferable to either the zero estimate from Horvitz-Thompson or the reduced database total of $\widetilde{T}_{y,\text{diff1}}$ from (2.6).

2.4 Multiple frames estimation under imperfect matching

2.4.1 Multiplicity difference estimator under imperfect matching

The results above assume one frame covers the universe, and this may not be possible in practice. In this section, we assume multiple frames cover the universe. Suppose the universe is partitioned into subpopulations U_g, and the subpopulations are divided into three groups, indexed by G_1, G_2, G_3:

$$U = \{\cup_{g \in G_1} U_g\} \cup \{\cup_{g \in G_2} U_g\}\{\cup_{g \in G_3} U_g\}.$$

If $g \in G_1$, U_g is covered by one or more frames, but not the database; if $g \in G_2$, U_g is covered by one or more frames and the database; if $g \in G_3$, U_g is covered only by the database. For the sample drawn from $\cup_{g \in G_1} U_g$ that is covered by one or more frames, we apply the Mecatti (2007) estimator to adjust for the multiple frames. Let s_{fg} denote the sample from frame f, and let $\widehat{T}_{fg} = \sum_{k \in s_{fg}} y_k / \pi_k^{(f)}$ denote the Horvitz-Thompson estimator for the sample from that frame. We define the coverage indicator as

$$F_{fg} = \begin{cases} 1, & \text{if subpopulation } U_g \text{ is covered by frame } f, \\ 0, & \text{otherwise.} \end{cases}$$

The weights that adjust for the multiplicity are

$$\psi_{fg} = \frac{F_{fg}}{\sum_f F_{fg}};$$

for example, if subpopulation U_g is covered by two frames f_1 and f_2, then $\psi_{f_1 g} = \psi_{f_2 g} = 1/2$, and $\psi_{fg} = 0$ for $f \neq f_1, f_2$. The unbiased multiplicity estimator for the $\sum_{g \in G_1} T_g$ is $\sum_{g \in G_1} \sum_{f=1}^{F} \psi_{fg} \widehat{T}_{fg}$.

For $g \in G_2$ under perfect matching, we could construct the multiplicity-based difference estimator,

$$\begin{aligned} \widetilde{T}_g^* &= \sum_{k \in U_g} \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{a}_\ell) + \sum_{f=1}^{F} \psi_{fg} \sum_{k \in s_{fg}} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{a}_\ell)}{\pi_k^{(f)}} \\ &= \sum_{k \in U_g} \widetilde{y}_k + \sum_{f=1}^{F} \psi_{fg} \sum_{k \in U_g} (y_k - \widetilde{y}_k) \frac{I_k^{(f)}}{\pi_k^{(f)}}. \end{aligned}$$

The unbiased difference estimator for $\sum_{g \in G_2} T_g$ is $\sum_{g \in G_2} \widetilde{T}_g^*$.

For $g \in G_3$, we can only predict with the auxiliary data since no frame covers this part of the population. Under perfect matching,

$$\widetilde{T}_g = \sum_{k \in U_g} \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{a}_\ell) = \sum_{k \in U_g} \widetilde{y}_k.$$

The synthetic predictor for $\sum_{g \in G_3} T_g$ is then $\sum_{g \in G_3} \widetilde{T}_g = \sum_{g \in G_3} \sum_{k \in U_g} \widetilde{y}_k$.

By combining the three groups, the multi-frame estimator under perfect matching is

$$\sum_{g \in G_1} \sum_{f=1}^{F} \psi_{fg} \widehat{T}_{fg} + \sum_{\ell \in \mathcal{A}} \left(\sum_{g \in G_2 \cup G_3} \sum_{k \in U_g} M_{k\ell} \right) \mu(\boldsymbol{a}_\ell) + \sum_{f=1}^{F} \psi_{fg} \sum_{k \in s_{fg}} \frac{y_k - \sum_{\ell \in \mathcal{A}} M_{k\ell} \mu(\boldsymbol{a}_\ell)}{\pi_k^{(f)}}. \quad (2.10)$$

If the matching is not perfect, replace $M_{k\ell}$ by match metrics $m_{k\ell}$, where $m_{k\ell} \in [0, 1]$. Because $m_{k\ell}$ is known only for $k \in s_{fg}$, we cannot just substitute $m_{k\ell}$ for $M_{k\ell}$ in the second term of (2.10) but the substitution is fine in the third term in (2.10). If ℓth record matches some element in $\cup_{g \in G_2 \cup G_3} U_g$, then $\left(\sum_{g \in G_2 \cup G_3} \sum_{k \in U_g} M_{k\ell} \right) = 1$. $\left(\sum_{g \in G_2 \cup G_3} \sum_{k \in U_g} M_{k\ell} \right)$ can be estimated from the sample, but we estimate $\left(\sum_{g \in G_2 \cup G_3} \sum_{k \in U_g} M_{k\ell} \right)$ as equal to 1 for simplicity. The multiplicity

adjusted difference estimator under imperfect matching is then

$$\widetilde{T}_{y,\text{diff,mult}} = \sum_{g\in G_1}\sum_{f=1}^{F}\psi_{fg}\widehat{T}_{fg} + \sum_{\ell\in\mathcal{A}}(1)\mu(\boldsymbol{a}_\ell) + \sum_{g\in G_2}\sum_{f=1}^{F}\psi_{fg}\sum_{k\in s_{fg}}\frac{y_k - m_{k\ell}\sum_{\ell\in\mathcal{A}}\mu(\boldsymbol{a}_\ell)}{\pi_k^{(f)}}.$$

The estimator is not unbiased and the bias depends on matching and prediction error. The expectation of the estimator is

$$\sum_{g\in G_1} T_g + \sum_{\ell\in\mathcal{A}}(1)\mu(\boldsymbol{a}_\ell) + \sum_{g\in G_2} T_g - \sum_{\ell\in\mathcal{A}}\left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\mu(\boldsymbol{a}_\ell).$$

The bias of the estimator is

$$\begin{aligned} & -\sum_{g\in G_3} T_g + \sum_{\ell\in\mathcal{A}}\mu(\boldsymbol{a}_\ell) - \sum_{\ell\in\mathcal{A}}\left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\mu(\boldsymbol{a}_\ell) \\ = \; & -\sum_{g\in G_3} T_g + \sum_{\ell\in\mathcal{A}}\mu(\boldsymbol{a}_\ell)\left[1 - \left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]. \end{aligned}$$

The estimator would be unbiased if $G_3 = \emptyset$ and $\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell} = 1$ for all $\ell \in \mathcal{A}$. Its variance by setting $m_{k\ell} \equiv 0$ for $k \in \cup_{g\in G_1} U_g$ is

$$\sum_{f=1}^{F}\sum_{g\in G_1\cup G_2}\sum_{g'\in G_1\cup G_2}\psi_{fg}\psi_{fg'}\sum_{j\in U_g}\sum_{k\in U'_g}\Delta_{jk}^{(f)}\frac{d_j}{\pi_j^{(f)}}\frac{d_k}{\pi_k^{(f)}},$$

where $d_j = y_j - \sum_{\ell\in\mathcal{A}} m_{j\ell}\mu(\boldsymbol{a}_\ell)$. If all $\pi_{jk}^{(f)} > 0$ in each frame, the unbiased variance estimator is given by

$$\sum_{f=1}^{F}\sum_{g\in G_1\cup G_2}\sum_{g'\in G_1\cup G_2}\psi_{fg}\psi_{fg'}\sum_{j\in U_g}\sum_{k\in U'_g}\Delta_{jk}^{(f)}\frac{d_j}{\pi_j^{(f)}}\frac{d_k}{\pi_k^{(f)}}\frac{I_j I_k}{\pi_{jk}^{(f)}}. \tag{2.11}$$

2.4.2 Properties of multiplicity difference estimator under imperfect matching

We describe some asymptotic properties of the multiplicity difference estimator here. The multiplicity adjusted difference estimator under imperfect matching (2.10) is asymptotically unbiased and design mean square consistent provided there is not too much matching error or undercoverage. To prove the results, we make the following assumptions,

(A1) As $N \to \infty$, $n^{(f)}N^{-1} \to \pi^{*(f)} \in (0,1)$. For all N, $\min_{j\in U} \pi_j^{(f)} \geqslant \lambda^{(f)} > 0$, and

$$\limsup_{N\to\infty} n^{(f)} \max_{j,k\in U: j\neq k} |\Delta_{jk}^{(f)}| < \infty.$$

(A2) The study variable $\{y_j\}_{j\in U}$ satisfy

$$\limsup_{N\to\infty} N^{-1} \sum_{j\in U} y_j^4 < \infty,$$

and the auxiliary information $\{\mu(\boldsymbol{a}_\ell)\}_{\ell\in\mathcal{A}}$ satisfy

$$\limsup_{N\to\infty} N^{-1} \sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)^4 < \infty.$$

(A3) $\sum_{g\in G_3} T_g^2 = O(N^\delta)$ with $\delta < 1$.

(A4) $\sum_{\ell\in\mathcal{A}} \left[1 - \left(\sum_{g\in G_2} \sum_{k\in U_g} m_{k\ell}\right)\right]^4 = O(N^\delta)$ with $\delta < 1$.

Remark. (A1) and (A2) are standard asymptotic assumptions in the survey literature (Breidt and Opsomer, 2000). (A3) ensures that there is not too much undercoverage, and (A4) ensures that there are not too much matching error.

Theorem 1. *Assume (A1)–(A4), then the multiplicity adjusted difference estimator is asymptotically design mean square consistent*

$$E\left[\left(\frac{\widetilde{T}_{y,\text{diff,mult}} - T_y}{N}\right)^2\right] \to 0 \text{ as } N \to \infty.$$

We prove the theorem by separately considering the bias and the variance in the following lemmas.

Lemma 1. *Assume (A2), (A3), and (A4), then the multiplicity adjusted difference estimator is asymptotically design unbiased*

$$E\left[\frac{\widetilde{T}_{y,\text{diff,mult}} - T_y}{N}\right] \to 0 \text{ as } N \to \infty.$$

Proof. The squared bias is

$$\begin{aligned}
\text{Bias}\left[\frac{\widetilde{T}_{y,\text{diff,mult}}}{N}\right]^2 &= \left[\frac{-\sum_{g\in G_3} T_g}{N} + \frac{\sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)\left[1 - \left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]}{N}\right]^2 \\
&= \left[\frac{-\sum_{g\in G_3} T_g}{N}\right]^2 + \left[\frac{\sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)\left[1 - \left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]}{N}\right]^2 \\
&+ 2\left[\frac{-\sum_{g\in G_3} T_g}{N}\right]\left[\frac{\sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)\left[1 - \left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]}{N}\right] \\
&= b_1 + b_2 + b_3
\end{aligned}$$

By (A3), $b_1 = \left[\left(-\sum_{g\in G_3} T_g\right)/N\right]^2 \leq (\sum_{g\in G_3} T_g^2)/N$, which goes to 0 as $N \to \infty$. By the Cauchy-Schwarz inequality and (A2), (A4),

$$\begin{aligned} b_2 &= \left[\frac{\sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)\left[1-\left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]}{N}\right]^2 \\ &\leq \frac{\sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)^2\left[1-\left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]^2}{N} \\ &\leq \left\{\frac{\sum_{\ell\in\mathcal{A}} \mu(\boldsymbol{a}_\ell)^4}{N}\right\}^{1/2}\left\{\frac{\sum_{\ell\in\mathcal{A}}\left[1-\left(\sum_{g\in G_2}\sum_{k\in U_g} m_{k\ell}\right)\right]^4}{N}\right\}^{1/2}, \end{aligned}$$

which goes to 0 as $N \to \infty$. Because cross product term b_3 goes to 0 by the Cauchy-Schwarz inequality, therefore $b_1 + b_2 + b_3$ goes to 0 as $N \to \infty$. □

Lemma 2. *Assume (A1) and (A2), then*

$$\begin{aligned} \mathit{Var}\left[N^{-1}\widetilde{T}_{y,\mathrm{diff,mult}}\right] &\leq \sum_{f=1}^{F}\frac{1}{N\lambda^{(f)}}\sum_{g\in G_1\cup G_2}\frac{\psi_{fg}\sum_{j\in U_g} d_j^2}{N} \\ &+ \sum_{f=1}^{F}\sum_{g\in G_1\cup G_2}\sum_{g'\in G_1\cup G_2}\psi_{fg}\psi_{fg'}\frac{\max_{j,k\in U_g: j\neq k}\left|\Delta_{jk}^{(f)}\right|}{\left(\lambda^{(f)}\right)^2}\left(\frac{\sum_{j\in U_g}|d_j|}{N}\right)^2 \to 0 \end{aligned}$$

as $N \to \infty$.

Theorem 1 then follows by Lemma 1 and Lemma 2.

2.5 Simulation experiment

2.5.1 Constructing the population and database

In the US state of South Carolina, there are about 500 operators of charter boats who take recreational angling trips with paying customers. Each boat can take multiple anglers, and over the course of 2016 there were about 50,000 angler trips on approximately 15,000 boat trips. These boat trips, along with the boat's logbook data on number of anglers and number of fish of each

species caught by those anglers, are required to be reported to the South Carolina Department of Natural Resources, though reporting is incomplete. After removing logbook reports with missing values, we took the remaining $N = 10,647$ as the universe U of actual boat trips to be studied. We then used a stochastic algorithm to simulate a corresponding database $\mathcal{A}$ of logbook records and a set of match metrics, $[m_{k\ell}]_{k\in U,\ell\in\mathcal{A}}$. In keeping with the real match metrics used in South Carolina, at most five of the $\{m_{k\ell}\}_{\ell\in\mathcal{A}}$ are non-zero for a given population element k.

We simulated the database by first sorting the universe in space and time, so that nearby elements in the population tend to be from the same coastal location and from nearby dates. We then used a Markov chain to determine the true (but unobservable) matching state of the population elements: no match, perfect match, high-quality match, or low-quality match. The transition

probability matrix of the chain is as follows:

State	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	ρ_0	ρ_1	ρ_2	0	0	0	0	ρ_3	0	0	0	0	0	0	0	0	0
1	ρ_0	ρ_1	ρ_2	0	0	0	0	ρ_3	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	ρ_0	ρ_1	ρ_2	0	0	0	0	ρ_3	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
16	ρ_0	ρ_1	ρ_2	0	0	0	0	ρ_3	0	0	0	0	0	0	0	0	0

where $\sum_{i=0}^{3} \rho_i = 1$. This chain determines that an element k has no match (state 0); or determines that element k has a perfect match (state 1); or determines that five successive elements $k, k+1, \ldots, k+4$ are high-quality (HQ) matches (states 2–6); or determines that ten successive elements $k, k+1, \ldots, k+9$ are low-quality (LQ) matches (states 7–16).

In the event of no match, no database record is created, and $m_{k\ell} = 0$ for all $\ell \in \mathcal{A}$.

In the event of a perfect match, a database record that matches element k is created, and $m_{k\ell} = 1$ for $k = \ell$ and zero otherwise.

In the event of five HQ matches, five database records are created: the first record matches element k, the next record matches element $k+1$, and so on until the fifth record matches element $k+4$. Further, we generate five match metric values that sum to one by independently generating $U_k, U_{k+1}, \ldots, U_{k+4}$ as Uniform(0,1) and setting

$$(m_{k+i,k}, m_{k+i,k+1}, \ldots, m_{k+i,k+4}) = \frac{1}{\sum_{i=0}^{4} U_{k+i}} (U_k, U_{k+1}, \ldots, U_{k+4})$$

and $m_{k+i,\ell} = 0$ otherwise for all five elements $i = 0, 1, \ldots, 4$. That is, all five elements have the same match metric values with the same five database records. If we sample one of these five elements, we know that it (in truth) matches one of the five database records with non-zero match metric values, but we do not know which one.

In the event of ten LQ matches, five database records are created: the first record matches element k, the next record matches element $k+1$, and so on until the fifth record matches element $k+4$. The remaining five elements have no matching database records. All ten population elements share the same match metric values, constructed similarly to those for the HQ matches, but with match metric values summing to $1/2$ instead of 1: independently generate $U_k, U_{k+1}, \ldots, U_{k+4}$ as Uniform(0,1) and set

$$(m_{k+i,k}, m_{k+i,k+1}, \ldots, m_{k+i,k+4}) = \frac{1}{2\sum_{i=0}^{4} U_{k+i}} (U_k, U_{k+1}, \ldots, U_{k+4})$$

and $m_{k+i,\ell} = 0$ otherwise for all ten elements $i = 0, 1, \ldots, 9$. Thus, if we sample one of these ten elements, we think there might be no match at all (true for half of the ten elements) or there might be a match among the five database records (true for half of the elements), but we do not know which one.

We consider two population/database combinations, determined by the choice of $\rho_0, \rho_1, \rho_2, \rho_3$. The "Poor Match" combination results in simulated proportions of match metric values that closely mirror those in the actual South Carolina data, while the "Better Match" combination has greatly improved matching:

Table 2.1: Summary for the choice of ρ_0, ρ_1, ρ_2, ρ_3 of Poor Match and Better Match.

	Records	ρ_0	ρ_1	ρ_2	ρ_3	No Match	LQ	HQ	Perfect
South Carolina						11.0%	52.5%	36.5%	0.0%
Poor Match	6836	0.35	0.20	0.25	0.20	8.6%	54.4%	31.7%	5.3%
Better Match	9031	0.10	0.20	0.60	0.10	2.3%	23.3%	69.8%	4.7%

Under the Poor Match combination, there are $6,836$ logbook records, so that many of the $N = 10,647$ population elements have no matching logbook records. Under the Better Match combination, there are $9,031$ logbook records. For each combination, we simulated the database once, and each population/database combination was then fixed for the remainder of the sampling experiment.

2.5.2 Estimation properties under repeated sampling for single frame

The sampling design used in our simulation study follows closely the design actually used by the Marine Recreational Information Program (MRIP) in South Carolina. We stratified the population into fifteen strata by crossing three regions (each consisting of contiguous South Carolina counties) and five waves (March–April, May–June, July–August, September–October, November–December). Similar to MRIP, our sampling design selects particular sites on particular days ("site-days") and intercepts all boat trips on those selected site-days. In MRIP, the site-days are selected with probability proportional to a measure of size that is an estimate of fishing activity ("pressure") for the site-day. In our design, we approximate this unequal probability design by allocating an overall sample size of $n = 500$ site-days to the 15 strata using a database estimate of fishing pressure for the stratum. We then selected site-days via simple random sampling without replacement within strata, and observed all boat-trips on selected site-days (there may, in fact, be no trips for a selected site-day). We chose $n = 500$ so that the number of selected site-days with non-zero fishing activity closely matches the 109 non-zero site-days for South Carolina in 2016. Site-days are thus the primary sampling units (PSUs), selected via stratified simple random sampling, and boat-trips are the secondary sampling units, selected with certainty within PSUs. Variance estimation needs to account for this stratified two-stage structure.

For each sampled boat-trip k in stratum h, the inclusion probability is $\pi_k = n_h/N_h$ where n_h is the number of site-days allocated to stratum h and N_h is the total number of site-days in stratum h, for $h = 1, 2, \ldots, 15$.

For this setting, our vector $\boldsymbol{a}_\ell$ of auxiliary information available for each element in the database includes time, location, number of anglers, and catch by species for multiple species of fish. Number of anglers and catch by species are of particular interest for estimation, and are observed for the sample of intercepted trips. The predictor $\mu(\boldsymbol{a}_\ell)$ for a characteristic of interest then simply returns the logbook value of the survey response: $\mu(\boldsymbol{a}_\ell) =$ logbook number of anglers when $y_k =$ intercepted number of anglers, $\mu(\boldsymbol{a}_\ell) =$ logbook number of black sea bass when $y_k =$ intercepted number of black sea bass, etc.

For each population/database combination, we drew 1000 independent stratified simple random samples from the fixed population and constructed the estimators $\widehat{T}_y$, $\widetilde{T}_{y,\text{diff1}}$, and $\widetilde{T}_{y,\text{diff2}}$ for several characteristics, including number of angler trips and total catch for red drum, black sea bass, gag grouper, Atlantic croaker, toadfish, and wahoo. These species were chosen to reflect a variety of reporting behaviors: in particular, they include species that are reported frequently in the database and are common enough to appear frequently in the on-site interviews, and species that are reported regularly but are rare enough to appear infrequently in the interviews. We also computed variance estimates as in (1.3), (2.5), and (2.9), but using the standard approximation of ignoring finite population corrections within strata. We present selected results here, noting that the Horvitz-Thompson estimator $\widehat{T}_y$ does not use the auxiliary information and has the same behavior under either combination.

Side-by-side boxplots for estimated total angler trips are shown in Figure 2.1, for estimated red drum catch in Figure 2.2, for estimated black sea bass catch in Figure 2.3, and for estimated gag grouper catch in Figure 2.4. We further summarized the results of the 1000 simulated samples for each estimator with the percent relative bias, root mean square error (RMSE), RMSE ratio (with Horvitz-Thompson estimator in the numerator), average estimated standard error (SE), and

coverage of nominal 95% confidence intervals computed assuming approximate normality. Results are presented in Table 2.2.

The Horvitz-Thompson estimator is theoretically unbiased, and both difference estimators are also nearly unbiased under each population/database combination and for each quantity of interest. Due to the low bias in all cases, the average estimated standard errors tend to be close to the RMSE's over the 1000 simulations, with the exceptions occurring for rarely-caught species like gag grouper. The sampling distributions of $\widehat{T}_y$ and $\widetilde{T}_{y,\text{diff2}}$, which are nonnegative by construction for nonnegative responses, are then highly skewed, with corresponding poor confidence interval coverage. The sampling distribution of $\widetilde{T}_{y,\text{diff1}}$, which is not constrained to be nonnegative, tends to be more symmetric and hence have better coverage with skewed distributions. This improved coverage comes at the expense of worse RMSE.

Under each population/database combination and for each quantity of interest, both difference estimators are better than the Horvitz-Thompson estimator, in terms of lower RMSE. The first difference estimator $\widetilde{T}_{y,\text{diff1}}$ is sometimes not much better than $\widehat{T}_y$, but the second difference estimator $\widetilde{T}_{y,\text{diff2}}$ is often much better than $\widehat{T}_y$, and is always better than $\widetilde{T}_{y,\text{diff1}}$.

2.5.3 Estimation properties under repeated sampling for multiple frame

In the multiple frame setting, we create two incomplete frames that together cover the universe and that partially overlap. We also consider two population/database combinations, "Poor Match" and "Better Match." The choices of ρ_0, ρ_1,ρ_2, ρ_3 are the same as in Table 2.1. The sampling design used in the simulation and the total sample size are the same as in the single frame. We choose $n = 250$ for each frame. We drew 1000 independent stratified samples and constructed $\widehat{T}_{y1}$, $\widehat{T}_{y2}$, $\widehat{T}_{y,\text{mec}}$, and $\widetilde{T}_{y,\text{diff,mult}}$ for several characteristics, where $\widehat{T}_{y1}$ is the Horvitz-Thompson estimator only from the first frame, $\widehat{T}_{y2}$ is the Horvitz-Thompson estimator only from the second frame, $\widehat{T}_{y,\text{mec}}$ is the Mecatti estimator adjusting for multiple frames but not using matched auxiliary information, and $\widetilde{T}_{y,\text{diff,mult}}$ uses information from both frames and the matched auxiliary information. We also compute the variance estimates as in (1.3), (1.8), and (2.11).

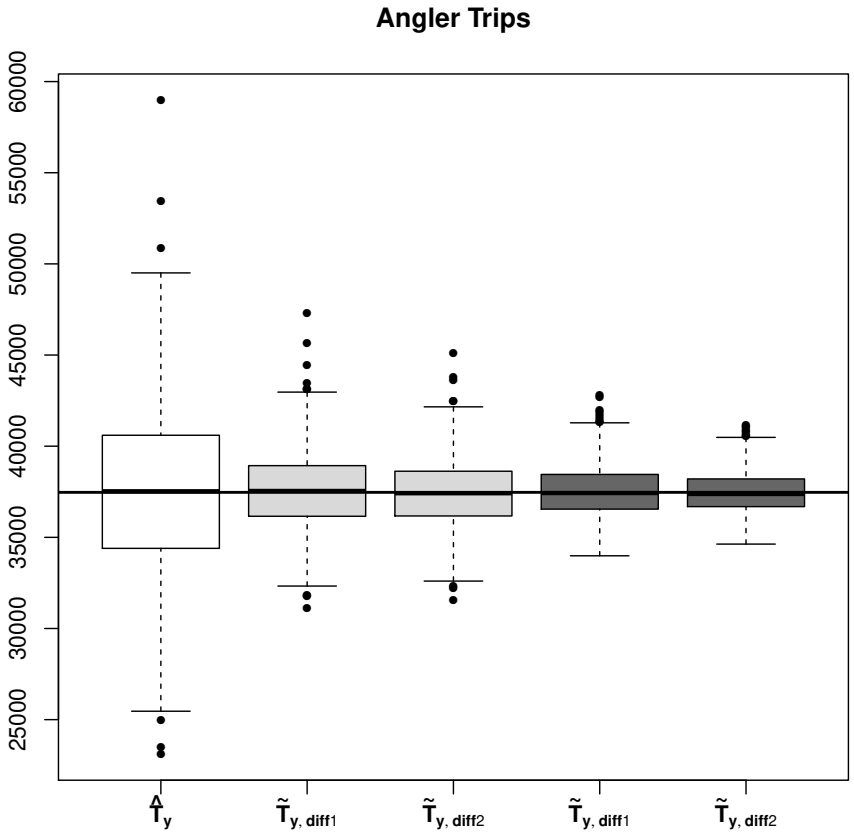

Figure 2.1: Boxplots for estimated total angler trips, based on 1000 simulated stratified simple random samples for each population/database combination. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_y$ (white boxplot) under either combination; $\widetilde{T}_{y,\mathrm{diff1}}$ and $\widetilde{T}_{y,\mathrm{diff2}}$ (light gray boxplots) under the Poor Match combination; $\widetilde{T}_{y,\mathrm{diff1}}$ and $\widetilde{T}_{y,\mathrm{diff2}}$ (dark gray boxplots) under the Better Match combination.

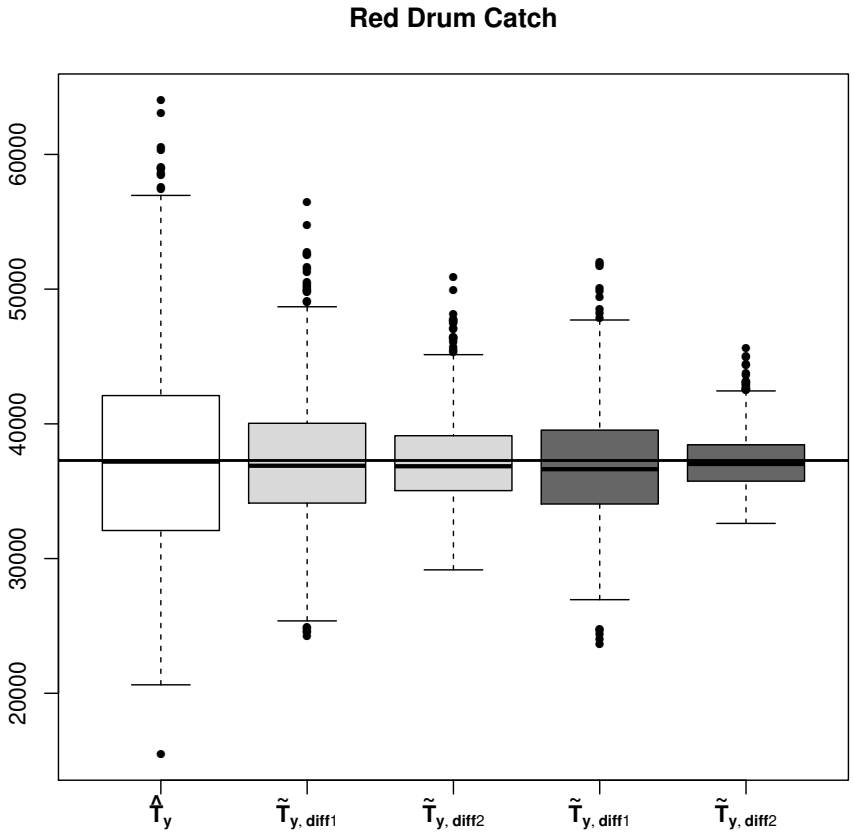

Figure 2.2: Boxplots for estimated total catch of red drum, based on 1000 simulated stratified simple random samples for each population/database combination. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_y$ (white boxplot) under either combination; $\widetilde{T}_{y,\text{diff1}}$ and $\widetilde{T}_{y,\text{diff2}}$ (light gray boxplots) under the Poor Match combination; $\widetilde{T}_{y,\text{diff1}}$ and $\widetilde{T}_{y,\text{diff2}}$ (dark gray boxplots) under the Better Match combination.

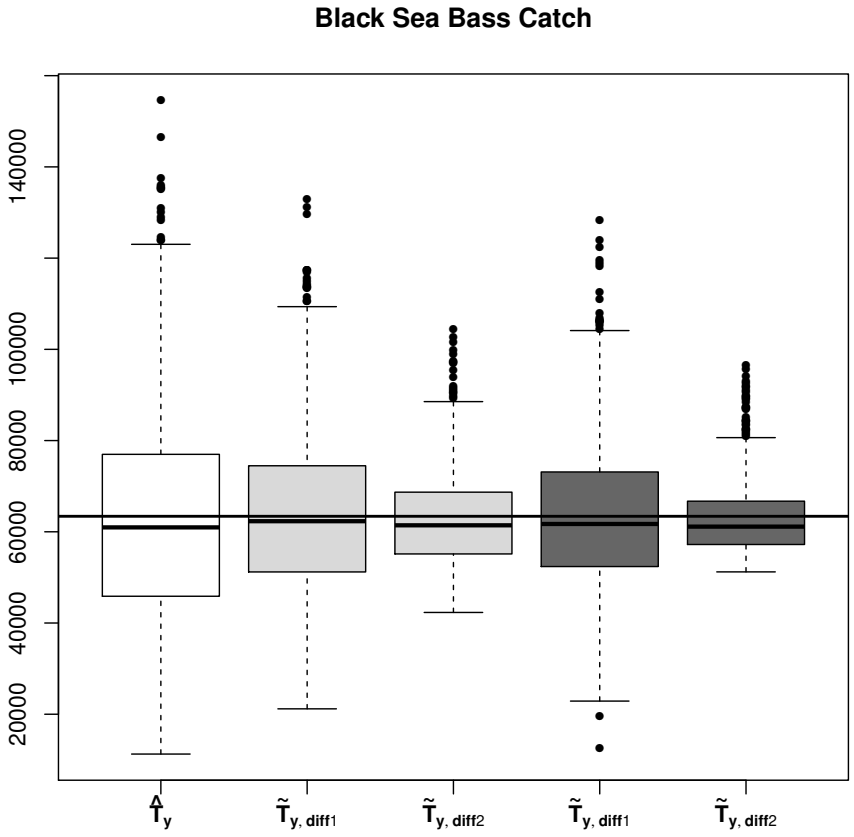

Figure 2.3: Boxplots for estimated total catch of black sea bass, based on 1000 simulated stratified simple random samples for each population/database combination. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_y$ (white boxplot) under either combination; $\widetilde{T}_{y,\text{diff1}}$ and $\widetilde{T}_{y,\text{diff2}}$ (light gray boxplots) under the Poor Match combination; $\widetilde{T}_{y,\text{diff1}}$ and $\widetilde{T}_{y,\text{diff2}}$ (dark gray boxplots) under the Better Match combination.

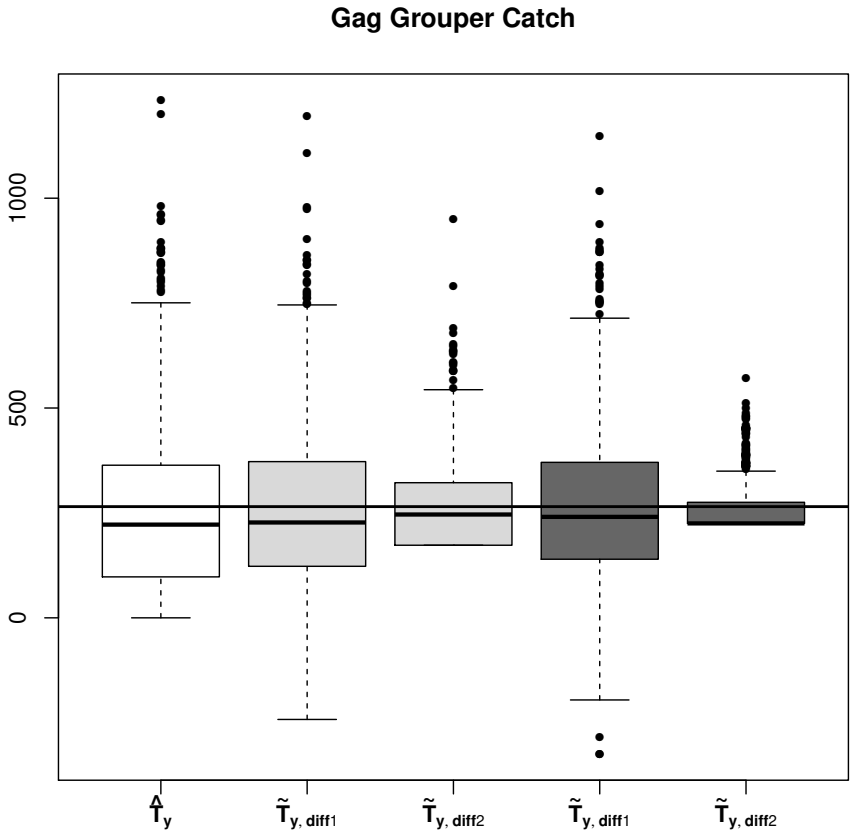

Figure 2.4: Boxplots for estimated total catch of gag grouper, based on 1000 simulated stratified simple random samples for each population/database combination. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_y$ (white boxplot) under either combination; $\widetilde{T}_{y,\text{diff1}}$ and $\widetilde{T}_{y,\text{diff2}}$ (light gray boxplots) under the Poor Match combination; $\widetilde{T}_{y,\text{diff1}}$ and $\widetilde{T}_{y,\text{diff2}}$ (dark gray boxplots) under the Better Match combination.

Table 2.2: Summary results for estimated angler trips, red drum catch, black sea bass catch, and gag grouper catch, based on 1000 simulated stratified simple random samples for each population/database combination. Relative RMSE (Root Mean Square Error) is RMSE of the estimator in the denominator and RMSE of $\widehat{T}_y$ in the numerator. Estimated SE (standard error) is for stratified simple random sampling, but ignoring within-stratum finite population corrections. Confidence interval coverage is for nominal 95% coverage under normality, using $(\text{estimator}) \pm 1.96 \times (\text{estimated SE})$.

			Poor Match		Better Match	
		$\widehat{T}_y$	$\widetilde{T}_{y,\text{diff1}}$	$\widetilde{T}_{y,\text{diff2}}$	$\widetilde{T}_{y,\text{diff1}}$	$\widetilde{T}_{y,\text{diff2}}$
Angler Trips	Mean	37567.7	37600.0	37424.6	37534.3	37471.0
	Percent Relative Bias	0.3	0.4	-0.1	0.2	0.0
	Relative RMSE	1.0	2.1	2.4	3.1	3.9
	RMSE	4427.2	2071.0	1828.9	1443.3	1138.3
	Average Estimated SE	4555.0	2155.3	1915.4	1476.5	1138.3
	Coverage	95.2	94.9	94.3	94.9	92.8
Red Drum Catch	Mean	37508.6	37266.3	37197.5	36887.2	37236.7
	Percent Relative Bias	0.6	0.0	-0.2	-1.1	-0.1
	Relative RMSE	1.0	1.5	2.3	1.7	3.5
	RMSE	7417.2	4857.0	3164.3	4301.2	2086.9
	Average Estimated SE	7270.7	4693.0	3042.5	4235.8	1960.8
	Coverage	93.1	92.6	90.5	92.8	87.6
Black Sea Bass Catch	Mean	63094.5	63915.1	62853.5	63509.8	62806.4
	Percent Relative Bias	-0.5	0.8	-0.9	0.2	-0.9
	Relative RMSE	1.0	1.3	2.2	1.4	3.0
	RMSE	23526.6	18008.0	10533.4	16287.8	7793.4
	Average Estimated SE	22725.9	17063.2	9653.7	15397.5	6473.7
	Coverage	87.5	91.2	83.1	92.4	76.9
Gag Grouper Catch	Mean	256.0	260.9	272.1	263.7	258.5
	Percent Relative Bias	-3.4	-1.5	2.7	-0.5	-2.5
	Relative RMSE	1.0	1.1	2.0	1.1	3.8
	RMSE	209.2	196.2	105.0	188.8	54.4
	Average Estimated SE	170.7	170.5	75.5	163.3	29.2
	Coverage	72.4	79.9	62.2	86.0	45.5

Figure 2.5 to Figure 2.8 are side-by-side boxplots for estimated angler trips, estimated red drum catch, estimated black sea bass catch, and estimated gag grouper catch. Table 2.3 summarizes the results for the Mecatti estimator and the multiplicity-adjusted difference estimator with the percent relative bias, RMSE, RMSE ratio (with Mecatti estimator in the numerator) average estimated standard error (SE), and coverage of nominal 95% confidence interval.

The figures show that each Horvitz-Thompson estimator is biased for the total because of the incomplete frames. The Mecatti estimator is unbiased, and the multiplicity-adjusted difference estimator is nearly unbiased. The multiplicity-adjusted difference estimator gains some information by using the auxiliary information and has lower RMSE than the Mecatti estimator. The Better Match difference estimator is better than the Poor Match difference estimator in terms of lower RMSE as can be seen from Table 2.3. The variance estimate is nearly unbiased and the confidence interval coverage is close to nominal 95%.

2.6 Discussion

The difference estimators described here are feasible in practice, given an auxiliary database and a suitable matching algorithm. The methodology offers substantial efficiency gains in a simulation study motivated by a real application in fisheries management. The simulation described here does not reflect any differential reporting, allowing probabilities of the match states to depend on the population characteristics. For example, boat captains catching only Atlantic croaker might be less likely to file a report than captains catching other species. The simulation also does not reflect differential measurement errors between the survey interviews and the logbook reports. In current practice, the boat captain is not required to file a logbook report immediately, and the catch recalled by the captain at the time of reporting may differ from the catch observed by an interviewer at a dockside intercept. These are directions for further study, both analytically and via simulation.

In results not reported here, we have also considered multiplicative adjustments of the Horvitz-Thompson estimator, as opposed to the additive adjustments of the difference-type estimators.

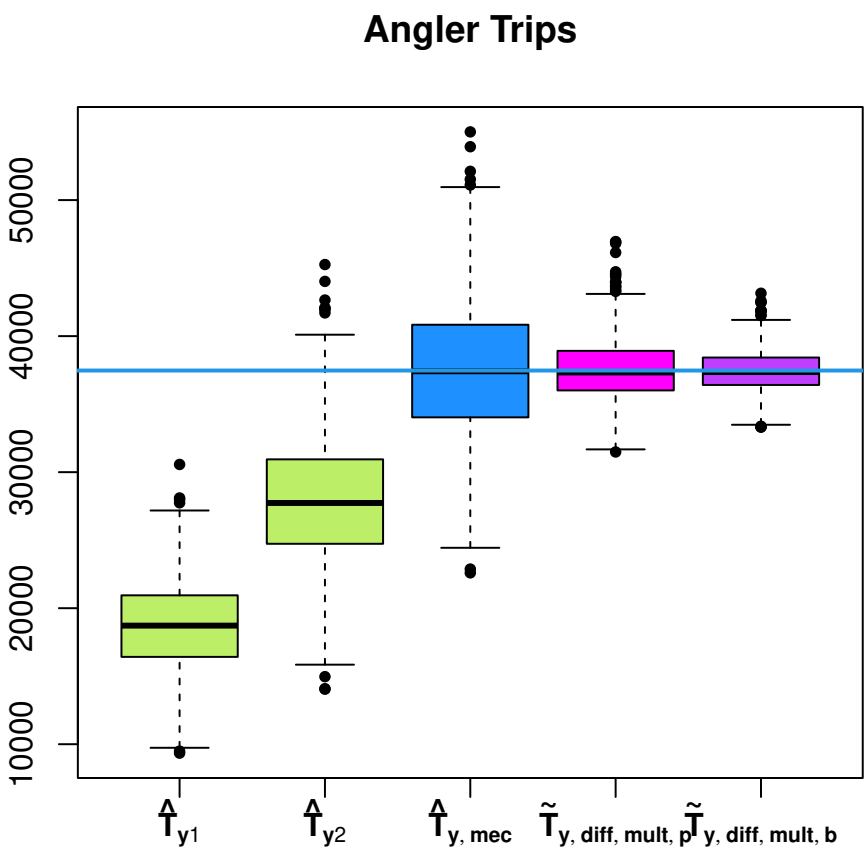

Figure 2.5: Boxplots for estimated total angler trips, based on 1000 simulated stratified simple random samples. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_{y1}$ and $\widehat{T}_{y2}$ (green boxplots) under either combination; Mecatti estimator $\widehat{T}_{y,\mathrm{mec}}$, (blue boxplot) under either combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (pink boxplot) under the Poor Match combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (purple boxplot) under the Better Match combination.

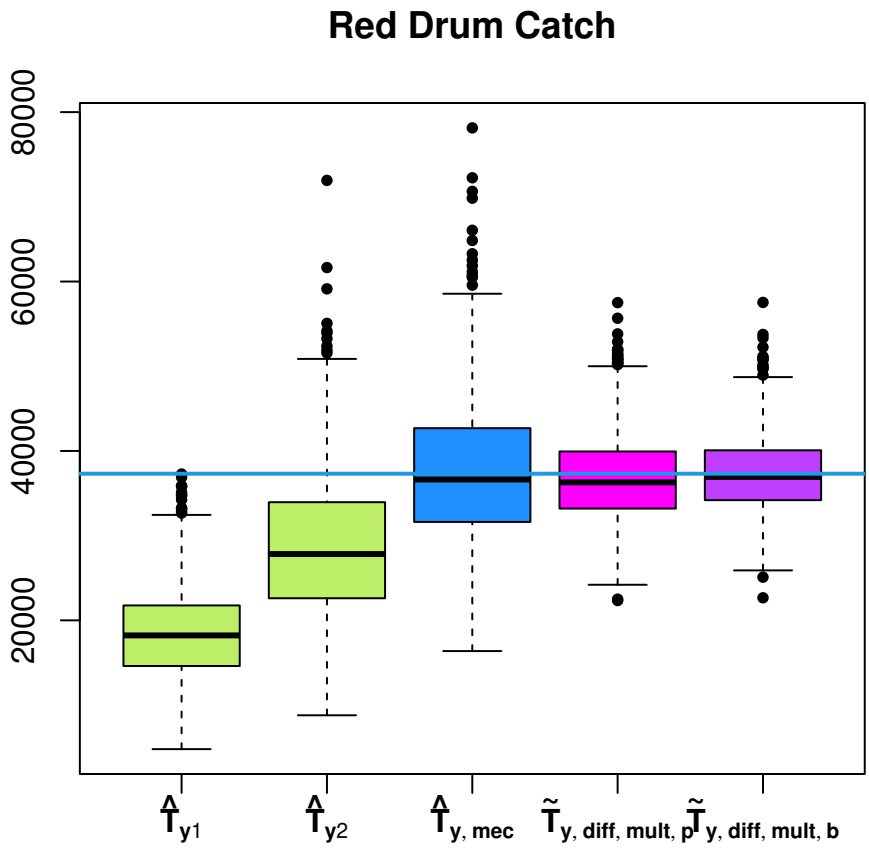

Figure 2.6: Boxplots for estimated total catch of red drum, based on 1000 simulated stratified simple random samples . Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_{y1}$ and $\widehat{T}_{y2}$ (green boxplots) under either combination; Mecatti estimator $\widehat{T}_{y,\mathrm{mec}}$, (blue boxplot) under either combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (pink boxplot) under the Poor Match combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (purple boxplot) under the Better Match combination.

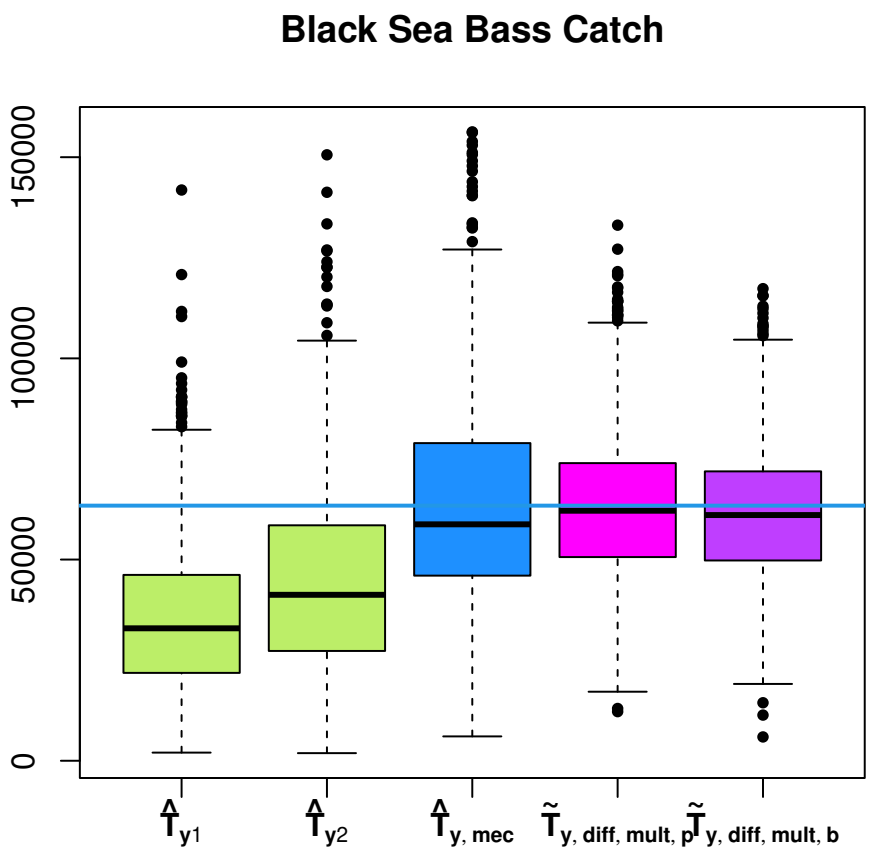

Figure 2.7: Boxplots for estimated total catch of black sea bass, based on 1000 simulated stratified simple random samples. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_{y1}$ and $\widehat{T}_{y2}$ (green boxplots) under either combination; Mecatti estimator $\widehat{T}_{y,\mathrm{mec}}$, (blue boxplot) under either combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (pink boxplot) under the Poor Match combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (purple boxplot) under the Better Match combination.

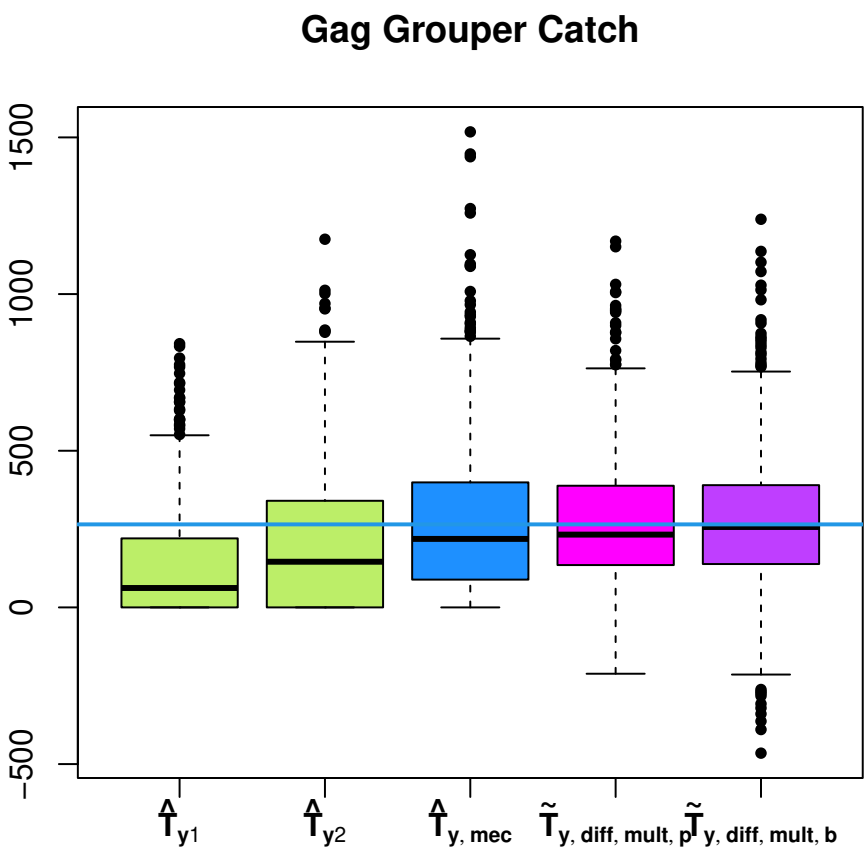

Figure 2.8: Boxplots for estimated total catch of gag grouper, based on 1000 simulated stratified simple random samples. Horizontal reference line is at the true value. From left to right: Horvitz-Thompson estimator $\widehat{T}_{y1}$ and $\widehat{T}_{y2}$ (green boxplots) under either combination; Mecatti estimator $\widehat{T}_{y,\mathrm{mec}}$, (blue boxplot) under either combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (pink boxplot) under the Poor Match combination; $\widetilde{T}_{y,\mathrm{diff,mult}}$ (purple boxplot) under the Better Match combination.

Table 2.3: Summary results for estimated angler trips, red drum catch, black sea bass catch, and gag grouper catch, based on 1000 simulated stratified simple random samples for each population/database combination. Relative RMSE (Root Mean Square Error) is RMSE of the estimator in the denominator and RMSE of $\widehat{T}_{y,\mathrm{mec}}$ in the numerator. Estimated SE (standard error) is for stratified simple random sampling, but ignoring within-stratum finite population corrections. Confidence interval coverage is for nominal 95% coverage under normality, using $(\text{estimator}) \pm 1.96 \times (\text{estimated SE})$.

		$\widehat{T}_{y,\mathrm{meca}}$	Poor Match $\widetilde{T}_{y,\mathrm{diff,mult}}$	Better Match $\widetilde{T}_{y,\mathrm{diff,mult}}$
Angler Trips	Mean	37656.4	37537.3	37423.7
	Percent Relative Bias	0.5	0.2	-0.1
	Relative RMSE	1.0	2.1	3.3
	RMSE	4987.5	2360.1	1514.5
	Average Estimated SE	4963.5	2298.4	1536.0
	Coverage	94.4	93.7	94.6
Red Drum Catch	Mean	37329.6	36871.0	37388.6
	Percent Relative Bias	0.1	-1.2	0.2
	Relative RMSE	1.0	1.5	1.8
	RMSE	8232.1	5432.3	4621.4
	Average Estimated SE	7863.8	5117.4	4430.1
	Coverage	90.7	90.4	94.7
Black Sea Bass Catch	Mean	63528.5	63551.4	61628.1
	Percent Relative Bias	0.2	0.2	-2.8
	Relative RMSE	1.0	1.4	1.5
	RMSE	2600.9	18532.1	16618.3
	Average Estimated SE	24485.7	17661.6	16169.2
	Coverage	86.5	91.5	92.8
Gag Grouper Catch	Mean	265.8	273.8	270.6
	Percent Relative Bias	0.3	3.3	2.1
	Relative RMSE	1.0	1.1	1.1
	RMSE	229.1	207	216.4
	Average Estimated SE	183.8	173.4	188.6
	Coverage	72.4	84.5	91.4

These multiplicative adjustments lead to ratio-type estimators that can be considered generalizations of capture-recapture sampling, extending the work of Liu et al. (2017). These multiplicative adjustments, however, seem particularly sensitive to poor matching and can have large biases and variances. Further study on such estimators is necessary.

We also proposed the difference estimator that allows for elements sampled from multiple frames with matching errors to the auxiliary database. Auxiliary information is still useful with multiple frames under imperfect matching. The variance estimator and confidence interval coverage work well in the multiple frame setting. A possible extension would be to include multiple auxiliary databases, and this will be more challenging if we consider the matching across multiple frames and matching across multiple auxiliary databases.

Chapter 3

Inference for complex surveys incorporating expert judgment at the screening stage

3.1 Introduction

Many complex surveys start with a "screener" stage: a probability sample of primary sampling units (PSUs) is drawn and the selected PSUs are examined to determine if they contain elements of interest. Any elements of interest are then either measured directly or subsampled further, depending on the context. Examples include household surveys looking for people in certain demographic groups (e.g., age-eligible children for an immunization survey), establishment surveys looking for certain specializations (e.g., hospitals with radiation oncologists), and area surveys looking for certain landscape characteristics (e.g., farms served by well water).

Our motivating example is a pilot study for the redesign of a recreational fisheries survey, the Large Pelagics Intercept Survey (LPIS), conducted by the US National Marine Fisheries Service (National Marine Fisheries Service, 2015). In this study, the PSUs are "site-days": saltwater fishing access sites (docks or marinas) on days of the year during the fishing season. A probability sample of site-days is selected via a stratified unequal-probability sampling design, and field crews visit the site-days to try to intercept boats returning from fishing trips that targeted pelagic species like tunas, billfishes, swordfish and sharks. LPIS data are used to estimate catch rate: average recreational catch per large pelagic boat trip, by species. Site-days with no large pelagic trips contribute nothing to estimation of catch rate. It is of interest to draw a site-day sample with high "yield," meaning as many trips as possible, so site-days are selected with probability proportional to their fishing pressure (an index of expected fishing activity).

Let $U = \{1, 2, \ldots, N\}$ denote the indices of the PSUs in the finite population of interest, let $\{z_k\}_{k \in U}$ denote the nonnegative integer-valued counts of the element of interest for each PSU,

and let $\{y_k\}_{k \in U}$ denote the PSU total of some characteristic for each PSU. In the above examples, k would index households, hospitals, land segments, or site-days; z_k would denote numbers of age-eligible children, radiation oncologists, farms with well water, or large pelagics trips; and y_k might denote numbers of immunized children, radiation oncology patients, wells with pesticide contamination, or catch of a given species. In each case, if $z_k = 0$ then $y_k = 0$. Our primary inferential target is the finite population rate $\phi_N = \sum_{k \in U} y_k / \sum_{k \in U} z_k = T_y / T_z$.

Let $s \subset U$ denote a sample of PSUs, selected via a sampling design $p(s)$ that assigns probability to all 2^N possible subsets of U, and assume (z_k, y_k) are observed without error for $k \in s$; we ignore subsampling within selected PSUs for simplicity of presentation. Define the sample membership indicators by $I_k = 1$ if $k \in s$ and $I_k = 0$ if $k \notin s$. Denote the first-order inclusion probabilities $\pi_k = \mathrm{P}(k \in s) = \mathrm{E}_p[I_k]$, where the probability is with respect to $p(\cdot)$. Provided $\pi_k > 0$ for all $k \in U$, the design is a probability sampling design, and the well-known HT estimators (Horvitz and Thompson, 1952) $\widehat{T}_z = \sum_{k \in U} z_k I_k \pi_k^{-1}$ and $\widehat{T}_y = \sum_{k \in U} y_k I_k \pi_k^{-1}$ are unbiased for the corresponding finite population totals T_z and T_y. The plug-in estimator $\widehat{\phi} = \widehat{T}_y / \widehat{T}_z$ is asymptotically unbiased and consistent for ϕ_N under mild conditions.

Intuitively, sampled elements with $z_k = y_k = 0$ contribute no information for estimation of the rate, ϕ_N. To see this more formally, consider a model ξ under which the y_k are uncorrelated, nonnegative random variables with means $\mathrm{E}_\xi[y_k \mid z_k] = \phi z_k$ and variances $\mathrm{Var}_\xi(y_k \mid z_k) = \sigma^2 z_k^{2\delta}$ for some $\phi, \delta, \sigma > 0$. Under this model, the plug-in estimator is unbiased for the finite population target: $\mathrm{E}_\xi[\hat{\phi} - \phi_N] = 0$. Further suppose that $z_k > 0$ for $k \in U^+$ and $z_k = 0$ for $k \in U \setminus U^+$ (so $y_k \equiv 0$ for $k \in U \setminus U^+$). Draw a sample $s = \{k \in U : I_k = 1\} \subset U$ via a probability sampling design on U, so that $\pi_k = \mathrm{E}_p[I_k] > 0$ for all $k \in U$. Under this design, the expected sample size is

$n = \sum_{k \in U} \pi_k$ and the plug-in estimator has anticipated variance (Isaki and Fuller, 1982)

$$\begin{aligned} \mathrm{E}_p \left[E_\xi \left[(\hat{\phi} - \phi_N)^2 \right] \right] &= \mathrm{E}_p \left[\frac{1}{\widehat{T}_z^2 T_z^2} \mathrm{Var}_\xi \left(\widehat{T}_y T_z - T_y \widehat{T}_z \right) \right] \\ &= \mathrm{E}_p \left[\frac{1}{\widehat{T}_z^2 T_z^2} \sum_{k \in U} \sigma^2 z_k^{2\delta} \left(\frac{T_z^2 I_k}{\pi_k^2} - \frac{2 T_z \widehat{T}_z I_k}{\pi_k} + \widehat{T}_z^2 \right) \right] \\ &\simeq \frac{\sigma^2}{T_z^2} \sum_{k \in U^+} z_k^{2\delta} \frac{1 - \pi_k}{\pi_k}, \end{aligned}$$

where the approximation arises by replacing unbiased estimators by their expectations in nonlinear functions.

Now suppose that a sample $s^+ = \{k \in U^+ : I_k^+ = 1\}$ could be allocated without error to U^+ only, with the probabilities increased to $\pi_k^+ = \mathrm{E}_p[I_k^+] = n\pi_k \left(\sum_{k \in U^+} \pi_k \right)^{-1} > \pi_k$ for all $k \in U^+$ (assuming for simplicity that these do not exceed 1, and rescaling otherwise). By the same argument above, the anticipated variance of the plug-in estimator is approximately

$$\frac{\sigma^2}{T_z^2} \sum_{k \in U^+} z_k^{2\delta} \frac{1 - \pi_k^+}{\pi_k^+} < \frac{\sigma^2}{T_z^2} \sum_{k \in U^+} z_k^{2\delta} \frac{1 - \pi_k}{\pi_k},$$

so that the design avoiding elements with zero responses is more efficient.

Perfect allocation to the nonzero part of the population is unlikely in practice. But in some surveys with a screener step, a sampler might be able to use expert judgment to improve the chances of finding PSUs that contain elements of interest. In our setting, local expertise might allow the field crew to improve chances of finding site-days with large pelagics trips. In other contexts, judgment might allow samplers to find more households with eligible children, or more establishments of the desired type. If samplers are allowed to use their judgment to find the samples, these samples are no longer probability samples, and making inference to the population of interest becomes more challenging.

We consider surveys in which the screener stage includes both a strict probability sample, selected with fixed and known inclusion probabilities, and an additional expert judgment sample

that we treat as a nonprobability sample. The presence of the probability sample makes it possible to estimate inclusion probabilities for the nonprobability sample and conduct valid inference to the finite population.

A growing literature discusses statistical inference for nonprobability samples. Ganesh et al. (2017) combine information from probability and nonprobability samples via small area estimation techniques, including bivariate Fay-Herriot models. Such approaches are not viable in our context due to the need to build separate models for each different characteristic. Sample matching approaches (Chen et al., 2020; Rivers, 2007; Yang et al., 2018) assign sampling weights from the probability sample to the nonprobability sample by comparing covariates available in both samples, and using the weight of the nearest probability sample element for each nonprobability sample element. These approaches could be implemented in our context but need to be modified since we have response variables observed in both the probability and nonprobability samples. The mass imputation approach (Chen et al., 2020; Kim et al., 2021) is a method that fits a model by regressing response variables on covariates in the nonprobability sample, then predicts ("imputes") the missing response variable using the observed covariates for every element in the probability sample. The estimator for the population is the weighted sum of the predicted values from the probability sample. The approach is also viable in our context but needs to be modified since we already have response variables in the probability sample. The approach of this paper follows the inverse weighting or quasi-randomization approach (e.g., Elliott and Valliant (2017), Chen et al. (2020), Kim and Wang (2019), and Valliant (2020)), in which the propensities for the nonprobability elements are estimated by combining the probability and nonprobability samples. In many of these papers, the response of interest is available only on the nonprobability sample, unlike our context. Doubly-robust estimation combines an estimated propensity and a regression model for response on covariates, so that inference is approximately unbiased if either model is correctly specified (Chen et al., 2020; Kim and Wang, 2019; Valliant, 2020). In our specific application, the weighted estimators for rates are doubly-robust under the model ξ above, and other covariates available for prediction of the responses are relatively weak, so that the doubly-robust estimators

were dominated by simpler inverse weighting estimators in our limited simulations. We restrict our discussion to inverse weighting estimators in the remainder of this chapter.

In section 3.2, we introduce notation and estimation of the inclusion probabilities for the nonprobability sample. Estimation for totals and rates is described in section 3.3, including "separate" and "combined" approaches to point estimation, both motivated by comparable estimators in dual-frame surveys. Variances of the estimators and variance estimators are derived. In section 3.4, we present some asymptotic properties of the combined estimator, assuming that the judgment selection follows Poisson sampling. In section 3.5, we describe simulation experiments using an artificial population constructed to mimic the LPIS motivating example. Across a range of conditions, the proposed combined strategy that allows for some judgment dominates the separate strategy and the classic strategy of pure probability sampling with known design weights. A brief discussion follows in section 3.6.

3.2 Sampling mechanisms and probability estimation

3.2.1 Probability and nonprobability sampling

Let s_0 denote a probability sample from U with known inclusion probabilities. A probability sample $s_A \subset s_0$ is selected with known inclusion probabilities and observations are obtained for $k \in s_A$. The elements $s_o \setminus s_A$ can either be left alone or moved via expert judgment to any other elements in $U \setminus s_A$. Whether moved or not, these selected elements constitute the nonprobability sample, $s_B \subset U \setminus s_A$, and $s_A \cap s_B = \emptyset$. The sample size of s_A is n_A, and the expected sample size of s_B is n_B, $n = n_A + n_B$. The probability sample indicators are I_k^A if $k \in s_A$, $I_k^A = 0$ otherwise; similarly, the nonprobability sample indicators are I_k^B if $k \in s_B$, $I_k^B = 0$ otherwise. The first-order inclusion probability for s_A is $\pi_k^A = \mathrm{E}\left[I_k^A\right] = \Pr\left[I_k^A = 1\right]$ satisfying $\pi_k^A > 0$ for all $k \in U$ and known for all $k \in s_A$. The first-order inclusion probability for the nonprobability sample is

$$\begin{aligned} \pi_k^B &= \Pr\left[k \in s_B \mid k \in s_A\right] \Pr\left[k \in s_A\right] + \Pr\left[k \in s_B \mid k \notin s_A\right] \Pr\left[k \notin s_A\right] \\ &= 0 + \rho_k(1 - \pi_k^A). \end{aligned} \tag{3.1}$$

Because of the judgment selection, the ρ_k and π_k^B are unknown and not necessarily positive for all $k \in U$.

3.2.2 Estimation of the inclusion probability for nonprobability samples

Estimation of the probabilities for the nonprobability sample requires modeling assumptions. We use an approach similar to Chen et al. (2020). For the kth element, we define $\boldsymbol{x}_k$ to be the vector of variables in the propensity model, possibly including y_k or z_k. We specify a parametric model, $f(\boldsymbol{x}_k, \boldsymbol{\theta})$, for ρ_k in (3.1), where $\boldsymbol{\theta}$ are the true unknown parameters. We estimate the parameters via a likelihood-based method. We assume Poisson sampling for s_B, under which the log-likelihood function is

$$\ln L(\boldsymbol{\theta}) = \sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho_k}{1-\rho_k}\right) + \sum_{k \in U \setminus s_A} \ln(1-\rho_k).$$

However, the second term of the log-likelihood involves data not in s_A or s_B. We thus replace the second term by the unbiased HT estimator (from the s_A sample) of its expectation, and compute the estimate $\widehat{\boldsymbol{\theta}}$ by maximizing the pseudo log-likelihood

$$\sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho_k}{1-\rho_k}\right) + \sum_{k \in U} \ln(1-\rho_k)(1-\pi_k^A)\frac{I_k^A}{\pi_k^A}.$$

We further assume a logistic model for ρ_k, $\text{logit}\,(\rho_k) = \boldsymbol{x}_k^\top \boldsymbol{\theta}$, for which the pseudo log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{k \in U \setminus s_A} I_k^B \boldsymbol{x}_k^\top \boldsymbol{\theta} - \sum_{k \in U} \ln\left\{1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta}\right)\right\}(1-\pi_k^A)\frac{I_k^A}{\pi_k^A},$$

and the score function is

$$\frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \sum_{k \in U \setminus s_A} I_k^B \boldsymbol{x}_k - \sum_{k \in U} \frac{\exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta}\right)}{1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta}\right)} \boldsymbol{x}_k \left(1-\pi_k^A\right)\frac{I_k^A}{\pi_k^A},$$

which has expected value equal to zero. The pseudo log-likelihood can be maximized by Newton-Raphson iteration or other numerical optimization. We plug in the estimated parameters $\widehat{\boldsymbol{\theta}}$ to obtain initial estimates, $\widetilde{\rho}_k$. A possible calibration approach would be to set the sample size equal to the

estimated expected sample size,

$$n_B = \sum_{k \in U} \widetilde{\rho}_k (1 - \pi_k^A), \tag{3.2}$$

but enforcing this calibration constraint is not feasible because the estimated $\widetilde{\rho}_k$ cannot be computed unless $\boldsymbol{x}_k$ are observed. Instead, we estimate the right-hand side of (3.2) from the probability sample by

$$\sum_{k \in s_A} \frac{\widetilde{\rho}_k (1 - \pi_k^A)}{\pi_k^A},$$

then ratio-adjust the initial estimates by finding the constant α such that

$$\left| n_B - \sum_{k \in s_A} \frac{\alpha \widetilde{\rho}_k (1 - \pi_k^A)}{\pi_k^A} \right|$$

is minimized, subject to the constraint that $\alpha \widetilde{\rho}_k \leq 1$. The initial $\widetilde{\rho}_k$ are then replaced by the ratio-adjusted values

$$\widehat{\rho}_k = \alpha \widetilde{\rho}_k.$$

We have also considered with-replacement sampling as an alternative model for the nonprobability sampling design. The pseudo log-likelihood and the score function under the with-replacement assumption are discussed in Appendix A.1. Simulation results reported below are similar under either the Poisson or with-replacement sampling model.

3.3 Estimation

The estimated inclusion probabilities for the nonprobability sample can be used to construct inverse probability weighting estimators, analogous to HT estimation. As in dual-frame estimation (see, for example, Singh and Mecatti (2011)), we could construct the estimator by keeping the probability sample and nonprobability sample separate or by combining them. We discuss both types of estimator and some properties, assuming known ρ_k, in the following subsections.

3.3.1 Separate estimator of the total

The separate estimator is a convex combination of the HT estimator from the probability sample s_A and the approximate HT estimator from the nonprobability sample s_B,

$$\widehat{T}_{y,\text{sep}} = \psi \sum_{k \in s_A} \frac{y_k}{\pi_k^A} + (1-\psi) \sum_{k \in s_B} \frac{y_k}{(1-\pi_k^A)\,\widehat{\rho}_k},$$

where ψ is between 0 and 1. The drawback of the separate estimator is that it might have large weights in the second term due to small estimated inclusion probabilities for the nonprobability sample.

Let $\widetilde{T}_{y,\text{sep}}$ denote the separate estimator with ρ_k replacing the estimates $\widehat{\rho}_k$. Then $\widetilde{T}_{y,\text{sep}}$ is exactly unbiased for T_y, and its design variance under Poisson sampling of s_B is

$$\text{Var}\left(\widetilde{T}_{y,\text{sep}}\right) = \sum_{k,\ell \in U} \sum \left(\pi_{k\ell}^A - \pi_k^A \pi_\ell^A\right) \frac{y_k^{**}}{\pi_k^A} \frac{y_\ell^{**}}{\pi_\ell^A} + (1-\psi)^2 \sum_{k \in U} \frac{(1-\rho_k)}{(1-\pi_k^A)\rho_k} y_k^2,$$

where $y_k^{**} = y_k \left\{\psi - (1-\psi)\pi_k^A (1-\pi_k^A)^{-1}\right\}$.

3.3.2 Combined estimator of the total

Denote the combined sample by $s = s_A \cup s_B$ and let $I_k = 1$ if $k \in s$ and $I_k = 0$ if $k \notin s$, and note that $I_k = I_k^A + (1 - I_k^A) I_k^B$. Then the combined (unconditional) first-order inclusion probability is

$$\pi_k = \Pr\left[k \in s\right] = \mathrm{E}\left[I_k\right] = \pi_k^A + (1-\pi_k^A)\rho_k.$$

By plugging in $\widehat{\rho}_k$ to estimate these unknown π_k, we obtain the combined estimator as the approximate HT estimator for the combined sample,

$$\widehat{T}_{y,\text{com}} = \sum_{k \in s} \frac{y_k}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k}.$$

Unlike the separate estimator, the combined estimator has stable weights by construction, since $\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k \geq \pi_k^A > 0$ and extremely small estimated inclusion probabilities are not possible.

Let $\widetilde{T}_{y,\mathrm{com}}$ denote the combined estimator with ρ_k replacing the estimates $\widehat{\rho}_k$. Then $\widetilde{T}_{y,\mathrm{com}}$ is exactly unbiased for T_y, and its design variance is

$$\mathrm{Var}\left[\widetilde{T}_{y,\mathrm{com}}\right] = \sum_{k,\ell \in U}\sum \mathrm{Cov}\left[I_k, I_\ell\right] \frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell} = \sum_{k,\ell \in U}\sum \left(\pi_{k\ell} - \pi_k\pi_\ell\right) \frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell},$$

where the second-order inclusion probability under Poisson sampling for s_B is

$$\begin{aligned} \pi_{k\ell} &= \mathrm{E}\left[I_k I_\ell\right] = \mathrm{E}\left[\left(I_k^A + (1 - I_k^A) I_k^B\right)\left(I_\ell^A + (1 - I_\ell^A) I_\ell^B\right)\right] \\ &= \pi_{k\ell}^A \left(1 - \rho_k\right)\left(1 - \rho_\ell\right) + \pi_k^A \rho_\ell \left(1 - \rho_k\right) + \pi_\ell^A \rho_k \left(1 - \rho_\ell\right) + \rho_k \rho_\ell. \end{aligned}$$

If s_A is selected via a measurable sampling design ($\pi_{k\ell}^A > 0$ for all $k, \ell \in U$), then the combined design is also measurable under Poisson sampling for s_B ($\pi_{k\ell} > 0$ for all $k, \ell \in U$ and any choice of $\{\rho_k\}$), so that an exactly unbiased variance estimator is

$$\widehat{V}_0\left(y\right) = \sum_{k,\ell \in U}\sum \left(\pi_{k\ell} - \pi_k\pi_\ell\right) \frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}\frac{I_k I_\ell}{\pi_{k\ell}}.$$

The standard approximate variance estimator, available in most statistical software, replaces the unbiased variance estimator $\widehat{V}_0$ by assuming a design with the same inverse-probability weights but with-replacement sampling,

$$\widehat{V}_1\left(y\right) = \frac{1}{n\left(n-1\right)} \sum_{k \in s} \left(\frac{y_k}{\pi_k / n} - \sum_{k \in s} \frac{y_k}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k} \right)^2. \tag{3.3}$$

The variance of the combined estimator can also be derived from the iterated variance formula by conditioning on the s_A sample. Assuming Poisson sampling for s_B, the variance is

$$\begin{aligned}
\text{Var}\left[\widetilde{T}_{y,\text{com}}\right] &= \text{Var}\left[\text{E}\left[\widetilde{T}_{y,\text{com}}\middle| s_A\right]\right] + \text{E}\left[\text{Var}\left[\widetilde{T}_{y,\text{com}}\middle| s_A\right]\right] \\
&= \sum_{k,\ell\in U}\left(\pi^A_{k\ell} - \pi^A_k\pi^A_\ell\right)\frac{y^*_k}{\pi^A_k}\frac{y^*_\ell}{\pi^A_\ell} + \text{E}\left[\text{Var}\left[\sum_{k\in U}\frac{(1-I^A_k)I^B_k}{\pi^A_k + (1-\pi^A_k)\rho_k}y_k\middle| s_A\right]\right] \quad (3.4) \\
&= \sum_{k,\ell\in U}\left(\pi^A_{k\ell} - \pi^A_k\pi^A_\ell\right)\frac{y^*_k}{\pi^A_k}\frac{y^*_\ell}{\pi^A_\ell} + \text{E}\left[\sum_{k\in U}\frac{(1-I^A_k)\rho_k(1-\rho_k)}{\left(\pi^A_k + (1-\pi^A_k)\rho_k\right)^2}y^2_k\right] \\
&= \sum_{k,\ell\in U}\left(\pi^A_{k\ell} - \pi^A_k\pi^A_\ell\right)\frac{y^*_k}{\pi^A_k}\frac{y^*_\ell}{\pi^A_\ell} + \sum_{k\in U}\frac{(1-\pi^A_k)\rho_k(1-\rho_k)}{\left(\pi^A_k + (1-\pi^A_k)\rho_k\right)^2}y^2_k, \quad (3.5)
\end{aligned}$$

where $y^*_k = (1-\rho_k)\pi^A_k y_k\left(\pi^A_k + (1-\pi^A_k)\rho_k\right)^{-1}$. This alternative expression for the variance suggests three additional approaches for variance estimation. The first approach, denoted as $\widehat{V}_2(y)$, estimates the first component of (3.5) from the probability sample s_A using the with-replacement sampling approximation, and estimates the expectation of conditional variance (the second component of (3.5)) with an unbiased estimator using the probability sample s_A only:

$$\begin{aligned}
\widehat{V}_2(y) &= \frac{1}{n_A(n_A-1)}\sum_{k\in s_A}\left(\frac{\widehat{y}^*_k}{\pi^A_k/n_A} - \sum_{k\in s_A}\frac{y_k(1-\widehat{\rho}_k)}{\pi^A_k + (1-\pi^A_k)\widehat{\rho}_k}\right)^2 \\
&+ \sum_{k\in U}\frac{(1-\pi^A_k)\widehat{\rho}_k(1-\widehat{\rho}_k)}{(\pi^A_k + (1-\pi^A_k)\widehat{\rho}_k)^2}y^2_k\frac{I^A_k}{\pi^A_k}, \quad (3.6)
\end{aligned}$$

where $\widehat{y}^*_k = (1-\widehat{\rho}_k)\pi^A_k y_k\left(\pi^A_k + (1-\pi^A_k)\widehat{\rho}_k\right)^{-1}$. The second approach, denoted as $\widehat{V}_3(y)$, estimates the first component of (3.4) from the probability sample s_A using the with-replacement sampling approximation, and estimates inside the expectation of (3.4), the conditional variance $\text{Var}\left(\widetilde{T}_{y,\text{com}}\middle| s_A\right)$ with the unbiased estimator using s_A and s_B, which is unbiased for the outer

expectation:

$$\begin{aligned}\widehat{V}_3(y) &= \frac{1}{n_A(n_A-1)}\sum_{k\in s_A}\left(\frac{\widehat{y}_k^*}{\pi_k^A/n_A}-\sum_{k\in s_A}\frac{y_k(1-\widehat{\rho}_k)}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}\right)^2 \\ &+ \sum_{k\in U}\frac{(1-I_k^A)I_k^B(1-\widehat{\rho}_k)}{\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)^2}y_k^2. \end{aligned} \tag{3.7}$$

The third approach, denoted as $\widehat{V}_4(y)$ estimates the first component of (3.4) from the probability sample s_A using with-replacement sampling approximation, and replaces the unbiased estimator of the second component used in $\widehat{V}_3(y)$ with its with-replacement approximation:

$$\begin{aligned}\widehat{V}_4(y) &= \frac{1}{n_A(n_A-1)}\sum_{k\in s_A}\left(\frac{\widehat{y}_k^*}{\pi_k^A/n_A}-\sum_{k\in s_A}\frac{y_k(1-\widehat{\rho}_k)}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}\right)^2 \\ &+ \frac{1}{n_B(n_B-1)}\sum_{k\in s_B}\left(\frac{(1-I_k^A)\widehat{\rho}_k n_B y_k}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}-\sum_{k\in s_B}\frac{y_k(1-I_k^A)}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}\right)^2. \end{aligned} \tag{3.8}$$

3.3.3 Estimation of rates

Since our primary target is the rate, we construct separate and combined versions of the plug-in rate estimator:

$$\widehat{R}_{\text{sep}}=\frac{\widehat{T}_{y,\text{sep}}}{\widehat{T}_{z,\text{sep}}}=\frac{\psi\sum_{k\in s_A}y_k/\pi_k^A+(1-\psi)\sum_{k\in s_B}y_k/\left(1-\pi_k^A\right)\widehat{\rho}_k}{\psi\sum_{k\in s_A}z_k/\pi_k^A+(1-\psi)\sum_{k\in s_B}z_k/\left(1-\pi_k^A\right)\widehat{\rho}_k},$$

and

$$\widehat{R}_{\text{com}}=\frac{\widehat{T}_{y,\text{com}}}{\widehat{T}_{z,\text{com}}}=\frac{\sum_{k\in s}y_k/\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)}{\sum_{k\in s}z_k/\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)}.$$

In simulation experiments described below, the combined estimator dominates the separate estimator across a range of conditions. Hence, we only discuss variance and variance estimation for the combined estimator of the ratio. Let $\widetilde{R}_{\text{com}}$ denote the rate estimator with ρ_k replacing the estimates $\widehat{\rho}_k$. The design variance of the ratio estimator can be approximated by Taylor expansion,

and the Taylor expansion of $\widetilde{R}_{\text{com}}$ is

$$\frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} + \sum_{k\in s} \frac{1}{(\pi_k^A + (1-\pi_k^A)\rho_k)} v_k,$$

where

$$v_k = \left\{ \frac{1}{\sum_{k\in U} z_k} \left(y_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} z_k \right) \right\}.$$

Then the design variance and the variance estimator of the ratio are

$$\text{Var}\left[\widetilde{R}_{\text{com}}\right] = \sum_{k,\ell\in U}\sum (\pi_{k\ell} - \pi_k\pi_\ell) \frac{v_k}{\pi_k}\frac{v_\ell}{\pi_\ell}; \quad \widehat{V}\left[\widetilde{R}_{\text{com}}\right] = \sum_{k,\ell\in U}\sum (\pi_{k\ell} - \pi_k\pi_\ell) \frac{\widehat{v}_k}{\pi_k}\frac{\widehat{v}_\ell}{\pi_\ell}\frac{I_k I_\ell}{\pi_{k\ell}},$$

where

$$\widehat{v}_k = \left\{ \frac{1}{\sum_{k\in s_A} z_k/\pi_k^A} \left(y_k - \frac{\sum_{k\in s_A} y_k/\pi_k^A}{\sum_{k\in s_A} z_k/\pi_k^A} z_k \right) \right\}.$$

We refer to this approximate variance estimator as $\widehat{V}_0$, and this variance estimate is often computed in standard software as a variance estimate of the ratio. Similarly, we could obtain $\widehat{V}_1$ by replacing y_k with $\widehat{v}_k$ in equation (3.3); $\widehat{V}_2$ by replacing y_k with $\widehat{v}_k$ in equation (3.6); $\widehat{V}_3$ by replacing y_k with $\widehat{v}_k$ in equation (3.7), and $\widehat{V}_4$ by replacing y_k with $\widehat{v}_k$ in equation (3.8) to get the variance estimator for the rate. These variance estimates of the rate are explored in the simulation study.

3.4 Asymptotic properties of the combined estimator

In this section, we provide some theoretical support for the combined estimator with Poisson sampling of s_B and known probabilities ρ_k, using an asymptotic framework in which there is a sequence of finite populations U_N of size N, with $N \to \infty$. Let F_N be the Nth realized finite population. Assume a sequence of probability samples $s_{A,N} \subset U_N$ of size $n_{A,N}$ and nonprobability samples $s_{B,N} \subset U_N$ of size $n_{B,N}$ drawn according to designs $p_{A,N}(\cdot)$ and $p_{B,N}(\cdot)$, respectively. The subscript N will be suppressed in much of the notation that follows.

We assume the following regularity conditions.

(B1) For the s_A sample, as $N \to \infty$, $n_A N^{-1} \to \pi^* \in (0,1)$. For all N, $\min_{k\in U} \pi_k^A \geqslant \lambda > 0$, $\min_{k,\ell} \pi_{k\ell}^A \geqslant \lambda^* > 0$ and

$$\limsup_{N\to\infty} n_A \max_{k,\ell\in U:k\neq\ell} \left|\pi_{k\ell}^A - \pi_k^A \pi_\ell^A\right| < \infty.$$

(B2) The study variables $\{y_k\}_{k\in U}$ satisfy

$$\limsup_{N\to\infty} N^{-1} \sum_{k\in U} y_k^4 < \infty.$$

(B3) For the s_A sample,

$$\lim_{N\to\infty} n_A^2 \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1}^A - \pi_{k_1}^A\right)\left(I_{k_2}^A - \pi_{k_2}^A\right)\left(I_{k_3}^A - \pi_{k_3}^A\right)\left(I_{k_4}^A - \pi_{k_4}^A\right)\right]\right| < \infty$$

and

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)\left(I_{k_3}^A I_{k_4}^A - \pi_{k_3k_4}^A\right)\right]\right| = 0,$$

where $D_{t,N}$ denotes the set of all distinct t-tuples $(k_1, k_2, \ldots, k_t)$ from U_N.

(B4)

$$n_A^{-1} n_B \to c,$$

where c is a positive constant.

Lemma 3. *Assume (B1), (B2), (B4), and s_B follows Poisson sampling with $0 \leqslant \rho_k \leqslant 1$. Then $\min_{k\in U} \pi_k \geqslant \lambda > 0$, $\min_{k,\ell\in U} \pi_{k\ell} \geqslant \lambda^* > 0$, and*

$$\limsup_{N\to\infty} n \max_{k,\ell\in U:k\neq\ell} \left|\pi_{k\ell} - \pi_k \pi_\ell\right| < \infty.$$

Lemma 4. *Assume (B1), (B2), (B3), (B4), and s_B follows Poisson sampling with $0 \leqslant \rho_k \leqslant 1$. Then*

$$\lim_{N\to\infty} n^2 \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|E\left[\left(I_{k_1}-\pi_{k_1}\right)\left(I_{k_2}-\pi_{k_2}\right)\left(I_{k_3}-\pi_{k_3}\right)\left(I_{k_4}-\pi_{k_4}\right)\right]\right| < \infty$$

and

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|E\left[\left(I_{k_1}I_{k_2}-\pi_{k_1k_2}\right)\left(I_{k_3}I_{k_4}-\pi_{k_3k_4}\right)\right]\right| = 0.$$

The proofs of the two lemmas above are provided in the Appendix A.2.

Result 1. *Assume (B1), (B2), (B4), and s_B follows Poisson sampling with $0 \leqslant \rho_k \leqslant 1$. Then*

$$Var\left[N^{-1}\widehat{T}_{y,\mathrm{com}}\right] \leq \frac{1}{N\lambda}\sum_{k\in U}\frac{y_k^2}{N} + \frac{\max_{k,\ell\in U:k\neq\ell}\left|\pi_{k\ell}-\pi_k\pi_\ell\right|}{\lambda^2}\left(\sum_{k\in U}\frac{|y_k|}{N}\right)^2 \to 0$$

as $N\to\infty$, hence the combined estimator is design mean square consistent.

Proof. The result follows from Lemma 3 and standard bounding arguments. □

Result 2. *Assume (B1), (B2), (B3), (B4), and s_B follows Poisson sampling with $0 \leqslant \rho_k \leqslant 1$. Then*

$$nE\left[\left|\widehat{V}\left[N^{-1}\widehat{T}_{y,\mathrm{com}}\right] - Var\left[N^{-1}\widehat{T}_{y,\mathrm{com}}\right]\right|\right] \to 0$$

as $N\to\infty$, so that $\widehat{V}\left[N^{-1}\widehat{T}_{y,\mathrm{com}}\right]$ is design consistent for Var $\left[N^{-1}\widehat{T}_{y,\mathrm{com}}\right]$.

Proof. The result follows from Lemma 3, Lemma 4, and Theorem 3 of Breidt and Opsomer (2000).

□

We next show that the combined estimator with known ρ_k is asymptotically normal, by adapting the argument of Theorem 1.3.6 in Fuller (2009). We assume a conventional finite population central limit theorem holds for $s_{A,N}$ to the finite population F_N, and assume nonprobability sample $s_{B,N}$ follows Poisson sampling, then the sequence of the combined estimator is asymptotically normal almost surely (a.s.).

Result 3. *Assume conditions (D1), (D2), (D3) of Appendix A.3, and s_B follows Poisson sampling with $0 \leqslant \rho_k \leqslant 1$. Then*

$$\left.\frac{\widehat{T}_{y,\mathrm{com}} - T_{y,N}}{\sqrt{V_A + V_B}}\right| F_N \xrightarrow{\mathcal{L}} N(0,1) \text{ a.s.},$$

where $T_{y,N} = \sum_{k \in U} y_k$, $V_A = \mathrm{Var}\left[E\left[\widehat{T}_{y,com}\middle| s_A, F_N\right]\middle| F_N\right]$,
and $V_B = E\left[\mathrm{Var}\left[\widehat{T}_{y,com}\middle| s_A, F_N\right]\middle| F_N\right]$. *Further,*

$$\frac{\widehat{T}_{y,\mathrm{com}} - \theta_N}{\sqrt{\nu_{0,N} + V_A + V_B}} \xrightarrow{\mathcal{L}} N(0,1),$$

where θ_N, $\nu_{0,N}$ is a sequence of constants not depending on F_N.

Proof. The proof is given in Appendix A.3. □

3.5 Simulation experiment

3.5.1 Constructing an artificial population

In the Large Pelagics Intercept Survey (LPIS), there are 57,388 site-days in the frame, comprising five months (June to October) and three states (Delaware, Massachusetts, Maryland). Each site-day has a value of "pressure" or expected fishing activity. We simulated the number of boat trips for each site-day as independent zero-inflated Poisson random variables, with parameters estimated from past LPIS data, and with values truncated at the maximum observed number of trips:

$$Z_k \sim \begin{cases} 0 & \text{with probability } p_k \\ \mathrm{Poisson}(\lambda_k) & \text{with probability } 1 - p_k, \text{ truncated at } \eta, \end{cases} \tag{3.9}$$

with $\mathrm{logit}(p_k) = \beta_0 + \beta_1 \times \mathrm{pressure}_k$, $\lambda_k = \alpha \times \mathrm{pressure}_k$. Parameter values vary from stratum to stratum; see the Appendix A.4 Table A.1 for details.

For each simulated trip, we then simulated the number of fish caught using eleven different parameterizations of the following model for catch, given trips:

$$Y_k \mid Z_k \sim \begin{cases} 0 & Z_k = 0, \\ \text{Binomial}(Z_k, \theta) & Z_k > 0, (\text{binary}) \\ (1-q) \times 0 + q\text{Poisson}(\theta Z_k^a) & Z_k > 0, \text{truncated at } Z_k\gamma, (\text{other catch}). \end{cases} \tag{3.10}$$

Catch is simulated independently across site-days. The parameter choices summarized in Table A.2 of Appendix A.4 reflect a variety of fishing behaviors and relationships between catch and trips: (1) *no relation* between catch and trips ($a = 0$); (2) *binary* with at most one fish caught per trip; (3,4,5) *retention* corresponding to a limit ($\gamma = 4$) on the number of fish, and expected catch proportional to *square root* of trips ($a = 0.5$), *linear* function of trips ($a = 1$), or *quadratic* function of trips ($a = 2$); (6,7,8) *moderate* catch ($\theta = 4$) with no limit ($\gamma = \infty$) and *square root*, *linear* or *quadratic* relation with trips; and (9,10,11) *high* catch ($\theta = 8$) with no limit ($\gamma = \infty$) and *square root*, *linear* or *quadratic* relation with trips. For the *high* catch scenarios, the species is not caught frequently (zero inflation with probability $1 - q = 0.6$), but is caught in high numbers when it is caught.

Trips and catch per trip are simulated once, then this fixed finite population is used in all subsequent sampling experiments.

3.5.2 Simulated samples

The sampling design used in the simulation follows the LPIS pilot study. We stratified the population into thirty strata by five months (June to October), three states (Delaware, Massachusetts, Maryland), and two kinds of boat mode (charter, private). The stratified sampling design selects particular site-days with probability proportional to fishing pressure. We allocate the overall sample size of 865 as in the actual pilot survey, then use the traditional LPIS sampling design to select a stratified unequal-probability sample, $s_o = \cup_{h=1}^{H} s_{oh}$, with probabilities proportional to fishing pressure.

Within each stratum h, we then divide s_{oh} at random, with 75% selected as s_{Ah}, a strict and unmovable probability sample, and the remaining 25% as the movable sample. We have two different methods for moving the movable sample. The first method, called the *stratum method*, requires that moves remain strictly within the stratum. The second method, called the *bucket method*, allows moves within a "bucket" consisting of two strata, corresponding to both boat modes within the same month, state, and kind-of-day (weekday or weekend). That is, boat mode can be changed in the bucket method, giving greater flexibility that was desired by field crews in the LPIS application.

For both methods of movement of the movable samples, we consider nine behaviors that reflect the different judgment abilities of field staff in finding site-days with non-zero trips: (1) *no move*, so that $s_{Bh} = s_{oh} \setminus s_{Ah}$ is the original 25% of the initial sample; (2) *unskilled*, in which s_{Bh} is a simple random sample from the complement of the stratum probability sample, $U_h \setminus s_{Ah}$; (3) *expert jump*, which successfully avoids all zero-trip site-days without distorting the distribution of non-zero-trip site-days; (4) *skilled jump*, which reduces the number of zero-trip site-days, without distortion of non-zero; (5) *pure tilt*, which does not change the number of zero trips but increases the probability of more trips when there are non-zero trips; (6) *jump and tilt*, which changes the distribution of both zero-trip and non-zero-trip site-days; (7) *skilled shift*, which is a special case of jump and tilt that leaves half the movable sample unmoved and moves the other half to the highest-trip site-days; (8) *logistic*, which creates inclusion probabilities as a function of trips using equation (3.11) and then draws s_{Bh} as a without-replacement sample from $U_h \setminus s_{Ah}$ using (approximately) these unequal probabilities; and (9) *with-replacement* that uses the same probabilities as (8) to draw s_{Bh} as a with-replacement sample from $U_h \setminus s_{Ah}$ using (exactly) these unequal probabilities.

For any judgment behavior, the final sample is $s = \cup_{h=1}^{H}(s_{Ah} \cup s_{Bh})$. The *no move* behavior yields the original probability sample, $s = s_o$, for which we can compute the standard survey-weighted estimator using the known sampling weights. This classic design/estimator strategy serves as the baseline for other strategies.

For all behaviors (including *no move*), we estimate the conditional inclusion probabilities ρ_k using the likelihood approach in Section 3.2, for both the Poisson and with-replacement models of selection. In all cases, we use the model

$$\text{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}_{\{Z_k=0\}} + \beta_2 Z_k \mathbf{1}_{\{Z_k>0\}} \tag{3.11}$$

so that probabilities depend on trips but not on other characteristics of the observed PSUs, including catch. This means estimated probabilities are the same for all catch characteristics, yielding weights that can be applied for all species. The model is misspecified for all of the judgment behaviors and is approximately correct (modulo normalization) for *logistic* and *with-replacement*. The initial estimate of ρ_k from the *stratum method* is calibrated to the stratum movable sample size n_{Bh}, and ρ_k from the *bucket method* is calibrated to the sum of movable sample size within a bucket of two strata. Because of the two different calibrations, estimators under the no move judgment behavior with the stratum movement method are different from estimators under the no move judgment behavior with the bucket movement method.

We draw 1000 stratified, unequal probability original samples from the population, use both movement methods and all nine judgment behaviors for each original sample, estimate ρ_k using both Poisson and with-replacement likelihoods and calibrate to the movable sample size, construct separate and combined estimators of catch rate using both sets of ρ_k estimates, and apply these design/estimator strategies to eleven species. The total number of strategies is

$$\begin{aligned}&(2 \text{ movement methods}) \times (9 \text{ judgment behaviors}) \times (2 \text{ likelihoods}) \times (2 \text{ estimators}) \\ &= 72 \text{ strategies},\end{aligned}$$

plus the baseline strategy with the no-move behavior and the original sample weights, for a total of 73 strategies.

We use all 73 strategies to estimate the catch rate for each of the 11 species in each of the 1000 samples. Figure 3.1 summarizes all results as boxplots of the root mean square error (RMSE) ratios

for the 11 estimated species catch rates. In the ratios, the RMSE of combined-Po, combined-WR, separate-Po or separate-WR is the numerator and the RMSE for no-move with original sample weights is the denominator. There are two boxplots for each judgment behavior and each estimator, with the left boxplot corresponding to the stratum movement method and the right boxplot corresponding to the bucket movement method. Figure 3.1 shows that strategies with the combined estimator (Poisson or with-replacement) dominate the baseline strategy across nearly all judgment behaviors and species characteristics. Strategies that use the separate estimator are not recommended: they sometimes beat the baseline strategy and sometimes are much worse. The combined strategies almost always dominate the separate strategies. Finally, the two combined strategies are very similar in most cases, so the choice of Poisson versus with-replacement likelihood is not very important. Between the two movement methods for movable samples, the bucket method is slightly better due to its greater flexibility.

In addition to (3.11), we have considered five other models: (1) $\text{logit}(\rho_k) = \beta_0 + \beta_1 \text{pressure}_k$; (2) $\text{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}_{\{Z_k=0\}}$; (3) $\text{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}_{\{Z_k=0\}} + \beta_2 Z_k \mathbf{1}_{\{Z_k>0\}} + \beta_3 Z_k^2 \mathbf{1}_{\{Z_k>0\}}$; (4) $\text{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}_{\{Z_k=0\}} + \beta_2 Z_k \mathbf{1}_{\{Z_k>0\}} + \beta_3 Z_k^2 \mathbf{1}_{\{Z_k>0\}} + \beta_4 X_k$; (5) $\text{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}_{\{Z_k=0\}} + \beta_2 Z_k \mathbf{1}_{\{Z_k>0\}} + \beta_3 Z_k^2 \mathbf{1}_{\{Z_k>0\}} + \beta_4 X_k$, where the X_k variable is the part of pressure_k that is orthogonal to trips, Z_k. We repeated the simulation experiment with all 72 strategies for these five different inclusion probability models, finding that models with trips dominate the model with pressure only, and that there is not much gained by adding complexity beyond model (3.11). The results are qualitatively similar and there is no specific model that dominates across judgment behaviors and 11 species. We therefore reported results only for model (3.11).

We compare variance estimators for the combined-Po estimator, which had the best overall performance among the point estimation strategies we considered. We consider $\widehat{V}_1$ by using the `survey` package in `R`, $\widehat{V}_2$, $\widehat{V}_3$, and $\widehat{V}_4$. We also considered replication methods for variance estimation, including a version of jackknife and of balanced repeated replication (BRR). Since neither performed as well as the simpler methods reported here, we omit the details and results for the replication methods. Boxplots of the 1000 estimators of non-replication methods for each of the

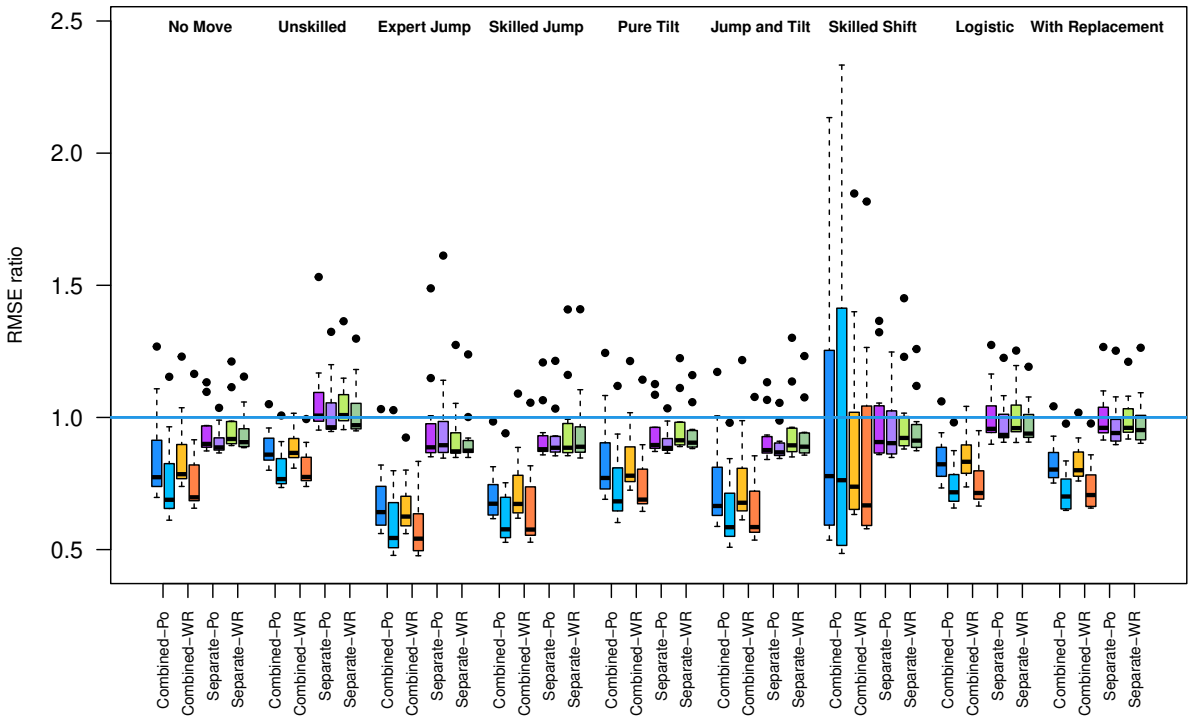

Figure 3.1: Ratios of RMSE for each strategy to RMSE of baseline strategy across 72 strategies and 11 species. Values greater than one favor the baseline strategy, which uses the no-move behavior and the original sample weights. Each pair of successive boxplots corresponds to RMSE ratios for one judgment behavior, one estimator type, one likelihood, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot). Combined-Po shows RMSE ratio boxplots under the two movement methods for the combined estimator with pseudo log-likelihood assuming Poisson sampling; Combined-WR is for the combined estimator with pseudo log-likelihood assuming with replacement sampling; Separate-Po is for the separate estimator with pseudo log-likelihood assuming Poisson sampling; and Separate-WR is for the separate estimator with pseudo log-likelihood assuming with replacement sampling.

11 species are provided in the supplementary materials. Figure 3.2 summarizes the results as boxplots of the relative root mean square error (RMSE) for variance estimate for the 11 species. In the ratios, RMSE of the estimated standard deviation is the numerator and the true standard deviation approximated via Monte Carlo is the denominator. There are two boxplots for each judgment behavior and each estimator, with the left boxplot corresponding to the stratum movement method and the right boxplot corresponding to the bucket movement method. The variance estimator $\widehat{V}_1$ is recommended since it is easily implemented in standard software and closer to the truth across the range of judgment behaviors, catch characteristics, and two moving methods. Figure 3.3 summarizes the coverage of the 95% confidence interval for the 11 species except for skilled shift judgment behavior, which is highly under-coverage due to the biased point estimate. The variance estimator $\widehat{V}_1$ also has better coverage across different catch characteristics and two moving methods.

3.6 Discussion

We propose estimators for inference when samplers are allowed to use judgment to select part of the samples. These estimators are based on simple (and incorrect) models, but work well in our simulations across a range of conditions. The combined estimator dominates the separate estimator and the classical probability estimator with known design weights across nearly all judgment behaviors. The simple variance estimator gives good confidence interval coverage. We also consider the behavior that moves all the movable samples to the highest trip site-days. The results are not included in Figure 3.1 because of more variation of RMSE ratio across 11 catch types that distort the structure of the figure.

Because of the problem of the possibly misspecified probability model, it is worth considering the doubly-robust estimation strategy. The strategy incorporates a prediction model for y_k. The estimator would remain consistent if at least one of the two models (the probability model or prediction model) is correct. In our settings, there is more than one way to do this, depending on which covariates we are going to use. We will discuss this in detail in chapter 4.

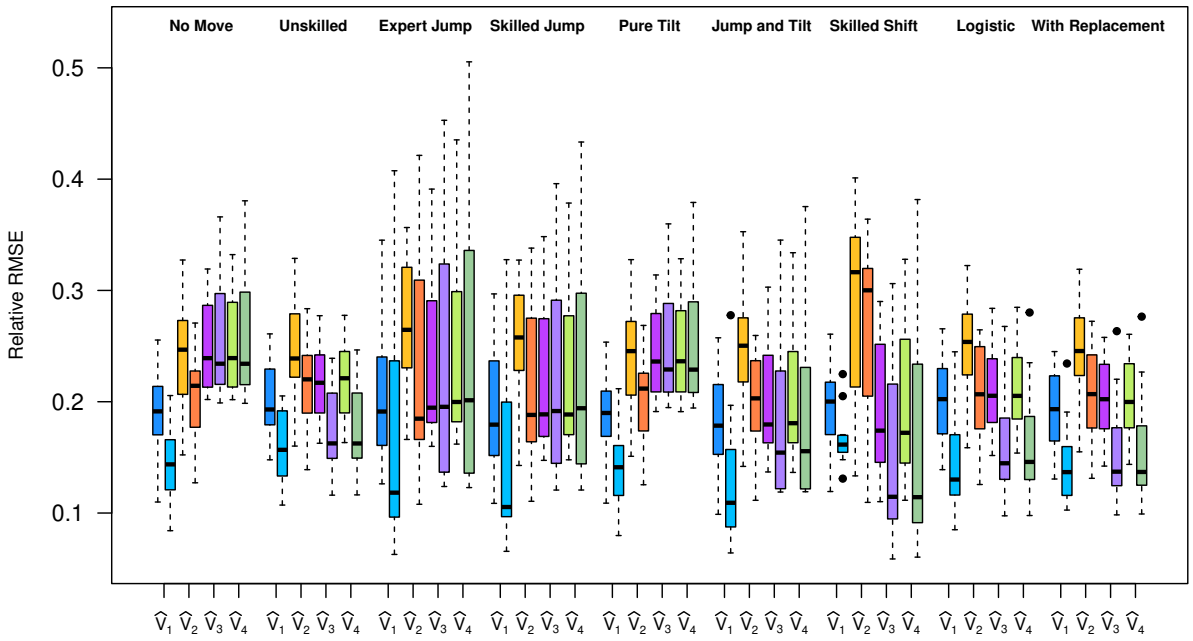

Figure 3.2: Relative RMSE of estimated standard deviation using four different variance estimators across 72 strategies and 11 species for the combined-Po estimator. Smaller values are better. Each pair of successive boxplots corresponds to relative RMSE for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

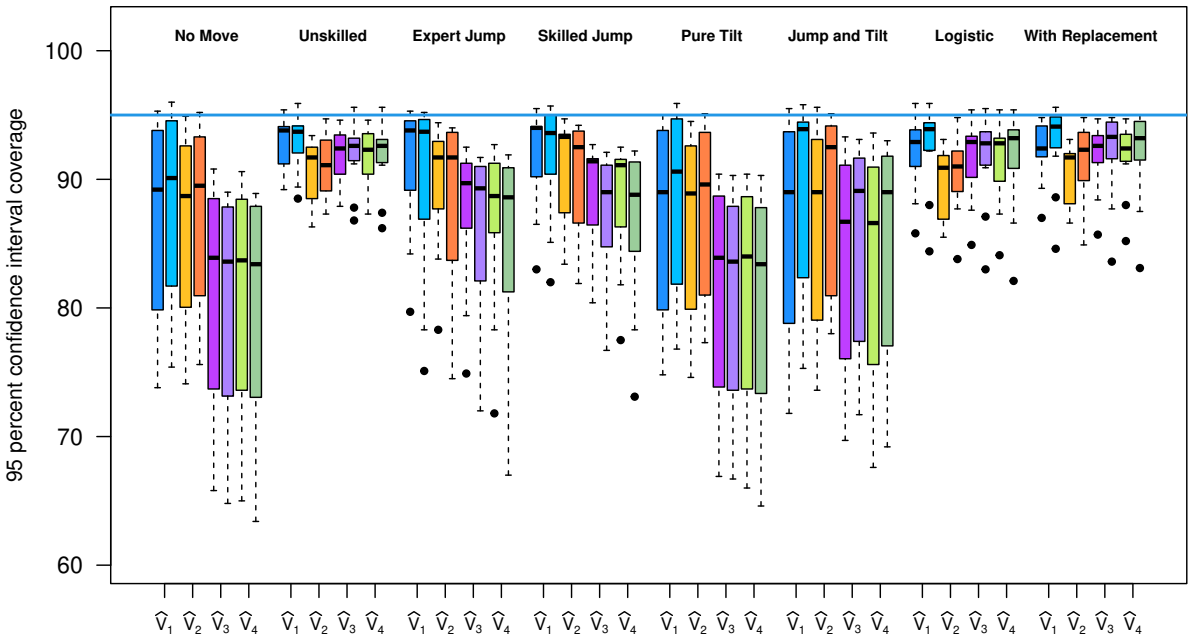

Figure 3.3: 95 percent confidence interval coverage across 64 strategies and 11 species using four different variance estimators. Each pair of successive boxplots corresponds to confidence interval coverage for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

The variance estimators considered here include only the design variance, ignoring the variation due to estimation of the parameters in the nonprobability model. In results not shown here, we have used the estimating equations approach to derive the variance that includes estimating the probabilities for the nonprobability sample. This approach only works if the probability model is correctly specified. In our simulations, the resulting variance estimate does not show a large difference with the variance estimate that only considered design. In practice, we recommend the Taylor approach since it is easy to implement and gives good confidence interval coverage across a range of conditions.

Chapter 4

Extension of inference for complex surveys incorporating expert judgment

4.1 Doubly-robust inference for complex surveys incorporating expert judgment

4.1.1 Introduction

In Chapter 3, we have proposed estimators that combine the strict probability sample and expert judgment sample by estimating the unknown inclusion probability of the judgment sample. Due to the possibly misspecified probability model, it is worth considering a prediction model and constructing a doubly-robust estimator, which would remain consistent if at least one of the two models (prediction model or probability model) is correctly specified.

Let $U = \{1, 2, \ldots, N\}$ denote the indices of the PSUs in the finite population of interest and our primary inferential targets are the finite population total $T_y = \sum_{k \in U} y_k$ and finite population rate $\phi_N = \sum_{k \in U} y_k / \sum_{k \in U} z_k = T_y / T_z$. Let $s \subset U$ denote a sample of PSUs, selected via a sampling design $p(s)$, and assume (z_k, y_k) are observed without error for $k \in s$. Let s_0 denote a probability sample from U with known inclusion probabilities. A probability sample $s_A \subset s_0$ is selected with known inclusion probabilities and observations are obtained for $k \in s_A$. The judgment sample is $s_B \subset U \setminus s_A$, where $s_A \cap s_B = \emptyset$ and $s = s_A \cup s_B$. The combined estimators for the total and rate are

$$\widehat{T}_{y,\text{com}} = \sum_{k \in s} \frac{y_k}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k},$$

and

$$\widehat{R}_{\text{com}} = \frac{\widehat{T}_{y,\text{com}}}{\widehat{T}_{z,\text{com}}} = \frac{\sum_{k \in s} y_k / \left(\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k\right)}{\sum_{k \in s} z_k / \left(\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k\right)},$$

where $\widehat{\rho}_k$ is the estimated inclusion probability for the judgment sample.

The combined estimator for rates is doubly-robust by construction. To see this formally, consider a model ξ under which the y_k are uncorrelated, nonnegative random variables with means $\mathrm{E}_\xi[y_k \mid z_k] = \phi z_k$ for some $\phi > 0$. The Taylor expansion of the rate is

$$\widehat{R}_{\text{com}} \approx \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} + \sum_{k\in s} \frac{1}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k} \left\{ \frac{1}{\sum_{k\in U} z_k} \left(y_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} z_k \right) \right\}.$$

If model ξ is correctly specified, we have

$$\begin{aligned} \mathrm{E}_\xi \left[\widehat{R}_{\text{com}} - \phi_N \right] &= \mathrm{E}_\xi \left[\mathrm{E}_\xi \left[\widehat{R}_{\text{com}} - \phi_N \middle| \{z_k\}_{k\in U} \right] \right] \\ &= \mathrm{E}_\xi \left[\frac{\phi \sum_{k\in s} z_k / \left(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k \right)}{\sum_{k\in s} z_k / \left(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k \right)} - \phi \right] \\ &= 0. \end{aligned}$$

On the other hand, if the probability model for ρ_k is correctly specified, we have

$$\mathrm{E}_p \left[\widehat{R}_{\text{com}} - \phi_N \right] \approx \sum_{k\in U} \left\{ \frac{1}{\sum_{k\in U} z_k} \left(y_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} z_k \right) \right\} \approx 0.$$

Under some mild assumptions, the rate estimator has the variance that goes to 0 under either the model ξ or the probability model. Because the bias goes to 0 if either the model ξ or the probability model is correctly specified, the estimated rate has the double robustness property.

There is some literature that discusses the doubly-robust inference with nonprobability samples. Chen et al. (2020) and Kim and Wang (2019) both build the doubly-robust estimator adapted from Kim and Haziza (2014) with nonprobability and probability samples to get the double robustness property. We also follow a similar approach by combining the probability and nonprobability samples. In section 4.1.2, we introduce the doubly-robust estimators of the dual-frame type and the variance of the estimators and variance estimators. In section 4.1.3, we describe simulation experiments of the doubly-robustness strategy.

4.1.2 Estimation

Doubly-robust estimator of the total

Suppose there is a parametric model for the response y, $\mathrm{E}[y|\,\boldsymbol{x}] = m(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}^\top \boldsymbol{\beta}$. We can construct the doubly robust estimator using the combined sample given by

$$\widehat{T}_{y,\mathrm{com,DR}} = \sum_{k \in s_A \cup s_B} \frac{y_k - m(\boldsymbol{x}_k, \widehat{\boldsymbol{\beta}})}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k} + \sum_{k \in U} m(\boldsymbol{x}_k, \widehat{\boldsymbol{\beta}}),$$

provided the auxiliary variables are known in the population. The coefficient of the model can be obtained from the ordinary least square or maximum likelihood from the combined sample. Let $\widetilde{T}_{y,\mathrm{com,DR}}$ denote the doubly-robust estimator with ρ_k replacing the estimates $\widehat{\rho}_k$, and $\boldsymbol{\beta}$ replacing the estimates $\widehat{\boldsymbol{\beta}}$. Then $\widetilde{T}_{y,\mathrm{com,DR}}$ is exactly unbiased for T_y. One way to get the design variance of the doubly-robust estimator $\widehat{T}_{y,\mathrm{com,DR}}$ is to approximate it by $\widetilde{T}_{y,\mathrm{com,DR}}$ and write the approximation as the weighted sum of y_k:

$$\begin{aligned}
& \widehat{T}_{y,\mathrm{com,DR}} \approx \widetilde{T}_{y,\mathrm{com,DR}} \\
= & \sum_{k \in s} y_k \left[\frac{1}{\pi_k^A + (1-\pi_k^A)\rho_k} \right. \\
- & \left. \left(\sum_{k \in s} \frac{\boldsymbol{x}_k^\top}{\pi_k^A + (1-\pi_k^A)\rho_k} - \sum_{k \in U} \boldsymbol{x}_k^\top \right) \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\pi_k^A + (1-\pi_k^A)\rho_k} \right)^{-1} \left(\frac{\boldsymbol{x}_k}{\pi_k^A + (1-\pi_k^A)\rho_k} \right) \right] \\
= & \sum_{k \in s} y_k w_k,
\end{aligned}$$

where

$$\begin{aligned}
w_k = & \left[\frac{1}{\pi_k^A + (1-\pi_k^A)\rho_k} \right. \\
- & \left. \left(\sum_{k \in s} \frac{\boldsymbol{x}_k^\top}{\pi_k^A + (1-\pi_k^A)\rho_k} - \sum_{k \in U} \boldsymbol{x}_k^\top \right) \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^\top}{\pi_k^A + (1-\pi_k^A)\rho_k} \right)^{-1} \left(\frac{\boldsymbol{x}_k}{\pi_k^A + (1-\pi_k^A)\rho_k} \right) \right].
\end{aligned}$$

The variance and variance estimator of the doubly-robust estimators are

$$\text{Var}\left[\widehat{T}_{y,\text{com,DR}}\right] \approx \sum_{k,\ell\in U}\sum \text{Cov}\left[I_k, I_\ell\right]\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell} = \sum_{k,\ell\in U}\sum \left(\pi_{k\ell}-\pi_k\pi_\ell\right)\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell},$$

and

$$\widehat{V}_0(y_{\text{DR}}) = \widehat{V}\left[\widehat{T}_{y,\text{com,DR}}\right] = \sum_{k,\ell\in U}\sum \left(\pi_{k\ell}-\pi_k\pi_\ell\right)\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}\frac{I_k I_\ell}{\pi_{k\ell}},$$

where π_k is $1/w_k$ and $\pi_{k\ell}$ is the second order inclusion probability of the corresponding weights in the estimators. The standard approximate variance estimator, available in most statistical software, replaces the unbiased variance estimator $\widehat{V}_0(y_{\text{DR}})$ by assuming a design with the same inverse-probability weights but with-replacement sampling,

$$\widehat{V}_1(y_{\text{DR}}) \;=\; \frac{1}{n\left(n-1\right)}\sum_{k\in s}\left(\frac{y_k}{\pi_k/n}-\sum_{k\in s}\frac{y_k}{\pi_k}\right)^2. \tag{4.1}$$

Similar to the combined estimator, the variance of the doubly-robust estimator can also be derived from the iterated variance formula by conditioning on the s_A sample. Assume Poisson sampling for s_B,

$$\begin{aligned}
&\text{Var}\left[\widetilde{T}_{y,\text{com,DR}}\right] \\
=\;& \text{Var}\left[\text{E}\left[\widetilde{T}_{y,\text{com,DR}}\middle| s_A\right]\right] + \text{E}\left[\text{Var}\left[\widetilde{T}_{y,\text{com,DR}}\middle| s_A\right]\right] \qquad (4.2)\\
=\;& \sum_{k,\ell\in U}\left(\pi^A_{k\ell}-\pi^A_k\pi^A_\ell\right)\frac{y^*_k}{\pi^A_k}\frac{y^*_\ell}{\pi^A_\ell} + \text{E}\left[\text{Var}\left[\sum_{k\in s}\frac{y_k-m(\boldsymbol{x}_k,\boldsymbol{\beta})}{\pi^A_k+(1-\pi^A_k)\rho_k}\middle| s_A\right]\right] \\
=\;& \sum_{k,\ell\in U}\left(\pi^A_{k\ell}-\pi^A_k\pi^A_\ell\right)\frac{y^*_k}{\pi^A_k}\frac{y^*_\ell}{\pi^A_\ell} + \sum_{k\in U}\frac{(1-\pi^A_k)\rho_k(1-\rho_k)}{\left(\pi^A_k+(1-\pi^A_k)\rho_k\right)^2}\left(y_k-\boldsymbol{x}_k^\top\boldsymbol{\beta}\right)^2 \qquad (4.3)
\end{aligned}$$

where $y^*_k = \left(y_k-\boldsymbol{x}_k^\top\boldsymbol{\beta}\right)(1-\rho_k)\pi^A_k\left(\pi^A_k+(1-\pi^A_k)\rho_k\right)^{-1}$. The expression suggests two different approaches for variance estimation. The first approach, denoted as $\widehat{V}_2(y_{\text{DR}})$, estimates the first component of (4.3) from the probability sample s_A using the with-replacement sampling approximation, and estimates the second component of (4.3) with an unbiased estimator using the

probability sample s_A only:

$$\begin{aligned}\widehat{V}_2(y_{\mathrm{DR}}) &= \frac{1}{n_A\,(n_A-1)}\sum_{k\in s_A}\left(\frac{\widehat{y}_k^*}{\pi_k^A/n_A}-\sum_{k\in s_A}\frac{\left(y_k-\boldsymbol{x}_k^\top\widehat{\boldsymbol{\beta}}\right)(1-\widehat{\rho}_k)}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}\right)^2 \\ &+ \sum_{k\in U}\frac{(1-\pi_k^A)\widehat{\rho}_k(1-\widehat{\rho}_k)}{(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k)^2}\left(y_k-\boldsymbol{x}_k^\top\widehat{\boldsymbol{\beta}}\right)^2\frac{I_k^A}{\pi_k^A}, \end{aligned} \tag{4.4}$$

where $\widehat{y}_k^* = \left(y_k-\boldsymbol{x}_k^\top\widehat{\boldsymbol{\beta}}\right)(1-\widehat{\rho}_k)\pi_k^A\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)^{-1}$. The second approach, denoted as $\widehat{V}_3\left(y_{\mathrm{DR}}\right)$, estimates the first component of (4.2) from the probability sample s_A using the with-replacement sampling approximation, and estimates inside the expectation of (4.2), the conditional variance $\mathrm{Var}\left(\widetilde{T}_{y,\mathrm{com,DR}}\middle|\, s_A\right)$ using s_A and s_B:

$$\begin{aligned}\widehat{V}_3\left(y_{\mathrm{DR}}\right) &= \frac{1}{n_A\,(n_A-1)}\sum_{k\in s_A}\left(\frac{\widehat{y}_k^*}{\pi_k^A/n_A}-\sum_{k\in s_A}\frac{\left(y_k-\boldsymbol{x}_k^\top\widehat{\boldsymbol{\beta}}\right)(1-\widehat{\rho}_k)}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}\right)^2 \\ &+ \sum_{k\in U}\frac{(1-I_k^A)^2I_k^B(1-\widehat{\rho}_k)}{\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)^2}\left(y_k-\boldsymbol{x}_k^\top\widehat{\boldsymbol{\beta}}\right)^2. \end{aligned} \tag{4.5}$$

Doubly-robust estimator of the rate

Although the combined estimator of the rate has the doubly robustness property under a linear model for y_k as a function of z_k, as shown in Section 4.1.1, we also consider a model-assisted doubly robust estimator to allow for more complex models. For example, the doubly-robust estimator of the rate that includes a quadratic term in the prediction model dominates the combined estimator in the simulation experiments described below. The doubly-robust estimator for the rate is

$$\widehat{R}_{\mathrm{com,DR}} = \frac{\widehat{T}_{y,\mathrm{com,DR}}}{\widehat{T}_{z,\mathrm{com,DR}}} = \frac{\sum_{k\in s_A\cup s_B}\frac{y_k-m_1(\boldsymbol{x}_k,\widehat{\boldsymbol{\beta}})}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}+\sum_{k\in U}m_1(\boldsymbol{x}_k,\widehat{\boldsymbol{\beta}})}{\sum_{k\in s_A\cup s_B}\frac{z_k-m_2(\boldsymbol{x}_k,\widehat{\boldsymbol{\beta}})}{\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k}+\sum_{k\in U}m_2(\boldsymbol{x}_k,\widehat{\boldsymbol{\beta}})}.$$

The design variance of the doubly-robust estimators for the rates could be approximated by Taylor expansion. For the weighted sum approach, the variance and variance estimator are

$$\text{Var}\left[\widehat{R}_{\text{com,DR}}\right] = \sum_{k,\ell\in U}\sum (\pi_{k\ell} - \pi_k\pi_\ell)\,\frac{v_k}{\pi_k}\frac{v_\ell}{\pi_\ell}; \quad \widehat{\text{V}}\left[\widehat{R}_{\text{com,DR}}\right] = \sum_{k,\ell\in U}\sum (\pi_{k\ell} - \pi_k\pi_\ell)\,\frac{\widehat{v}_k}{\pi_k}\frac{\widehat{v}_\ell}{\pi_\ell}\frac{I_k I_\ell}{\pi_{k\ell}},$$

where π_k is $1/w_k$,

$$v_k = \left\{\frac{1}{\sum_{k\in U} z_k}\left(y_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} z_k\right)\right\},$$

and

$$\widehat{v}_k = \left\{\frac{1}{\sum_{k\in s_A} z_k/\pi_k^A}\left(y_k - \frac{\sum_{k\in s_A} y_k/\pi_k^A}{\sum_{k\in s_A} z_k/\pi_k^A} z_k\right)\right\}.$$

We compute the variance estimator using the standard approximate variance estimator from with-replacement sampling, denoted as $\widehat{V}_1(v_{\text{DR}})$, which replaces equation (4.1) by $\widehat{v}_k$. For the iterated variance approach under Poisson sampling of s_B sample,

$$\begin{aligned}
&\text{Var}\left[\widehat{R}_{\text{com,DR}}\right] = \text{Var}\left[\text{E}\left[\widehat{R}_{\text{com,DR}}\middle| s_A\right]\right] + \text{E}\left[\text{Var}\left[\widehat{R}_{\text{com,DR}}\middle| s_A\right]\right] \qquad (4.6)\\
= &\sum_{k,\ell\in U}\left(\pi_{k\ell}^A - \pi_k^A\pi_\ell^A\right)\frac{v_k^*}{\pi_k^A}\frac{v_\ell^*}{\pi_\ell^A}\\
+ &\sum_{k\in U}\frac{(1-\pi_k^A)\rho_k(1-\rho_k)}{(\pi_k^A + (1-\pi_k^A)\rho_k)^2}\left\{\frac{1}{\sum_{k\in U} z_k}\left((y_k - m_1(\boldsymbol{x}_k,\boldsymbol{\beta})) - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k}(z_k - m_2(\boldsymbol{x}_k,\boldsymbol{\beta}))\right)\right\}^2,
\end{aligned}$$

where

$$v_k^* = \left\{\frac{1}{\sum_{k\in U} z_k}\left((y_k - m_1(\boldsymbol{x}_k,\boldsymbol{\beta})) - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k}(z_k - m_2(\boldsymbol{x}_k,\boldsymbol{\beta}))\right)\right\}\frac{(1-\rho_k)\pi_k^A}{(\pi_k^A + (1-\pi_k^A)\rho_k)}.$$

We estimate the first component of (4.6) from the probability sample s_A using the with-replacement sampling approximation, and estimate the second term of (4.6) either from the expectation of conditional variance using s_A only or estimate the conditional variance with the unbiased estimator using s_A and s_B. The estimators are denoted as $\widehat{V}_2(v_{\text{DR}})$ and $\widehat{V}_3(v_{\text{DR}})$.

The estimators above consider only the design variance, and ignore the variation caused by the estimation of the parameters. Except for the design variance, we have used the estimating equations approach to derive the variance that includes the estimation of the parameters in the nonprobability and prediction model. However, this more complex estimator does not dominate the estimators above in our simulation experiments. The variance estimator in Chen et al. (2020) implemented from Kim and Haziza (2014) is not applicable in our setting, since we will generally have different dimensions for the prediction model and the probability model.

4.1.3 Simulation experiment

We use the same simulated population from section 3.5 with the same stratified design and sample size. The proportion of the judgment samples, judgment behavior, and the methods for movable sample are the same as before. For all behaviors, we estimate the conditional inclusion probabilities ρ_k using the likelihood from the Poisson model of selection with covariates as indicators of trip,

$$\text{logit}(\rho_k) = \alpha_0 + \alpha_1 \mathbf{1}_{\{Z_k=0\}} + \alpha_2 Z_k \mathbf{1}_{\{Z_k>0\}}.$$

We consider two different prediction models in the simulation, one linear in trips and one quadratic in trips:

$$m(Z_k) = \beta_1 Z_k,$$

$$m(Z_k) = \beta_1 Z_k + \beta_2 Z_k^2$$

where trips are assumed available at the population level and the total trips is known. In general, this is not always a realistic assumption but in some cases, we might have good estimates of total trips from another external source, like a separate survey or monitoring program.

We draw 1000 stratified, unequal probability original samples from the population, use all nine judgment behaviors for each original sample, estimate ρ_k using Poisson likelihood, construct combined and doubly-robust estimators of catch and catch rate, and apply to eleven species for two

movement methods. The total number of strategies is

$$\begin{aligned}&(2 \text{ movement methods}) \times (9 \text{ judgment behaviors}) \times \\ &(2 \text{ prediction models of doubly-robust estimator} + \text{combined estimator}) = 54 \text{ strategies},\end{aligned}$$

plus the baseline strategy that uses the survey regression estimator for the original probability sample with known design weights. Therefore, there is a total of 55 strategies.

Figure 4.1 shows that the catch estimator is nearly unbiased if at least one of the two models is correctly specified: either catch is a linear (or quadratic) function of trips, or the judgment behavior is logistic or with-replacement. Figure 4.2 shows another example in which the catch estimator is nearly unbiased if at least one of the two models is correctly specified: either catch is a quadratic function of trips, or the judgment behavior is logistic or with-replacement. For other catch characteristics, the figures show results that are qualitatively similar; these figures are omitted. Results summarized across trips are presented in Figure 4.3, which shows all results as boxplots of the root mean square error (RMSE) ratios for estimated catch for the 11 species. In the ratios, combined-Po, DR-Trip, DR-Quad-Trip is the numerator and RMSE for no-move with original sample weights of regression estimator is the denominator. There are two boxplots for each judgment behavior and each estimator, with the left boxplot corresponding to the stratum movement method and the right boxplot corresponding to the bucket movement method. Figure 4.3 shows that the combined estimator and doubly-robust estimator using some judgment sampling dominate the regression estimator using pure probability sampling across nearly all judgment behaviors and species.

Figure 4.4 to Figure 4.7 summarize the results of the bias ratio and root mean square error (RMSE) ratios for the estimated catch for all 11 species. Figure 4.4 and Figure 4.5 are the summary of ratios when the prediction model is linear in trips. Figure 4.7 is the summary of ratios when the prediction model is quadratic in trips. In the ratios, the absolute bias or RMSE for the doubly-robust estimator is the numerator and absolute bias or RMSE for the combined estimator is the denominator. These figures show that doubly-robust estimator with two different prediction models

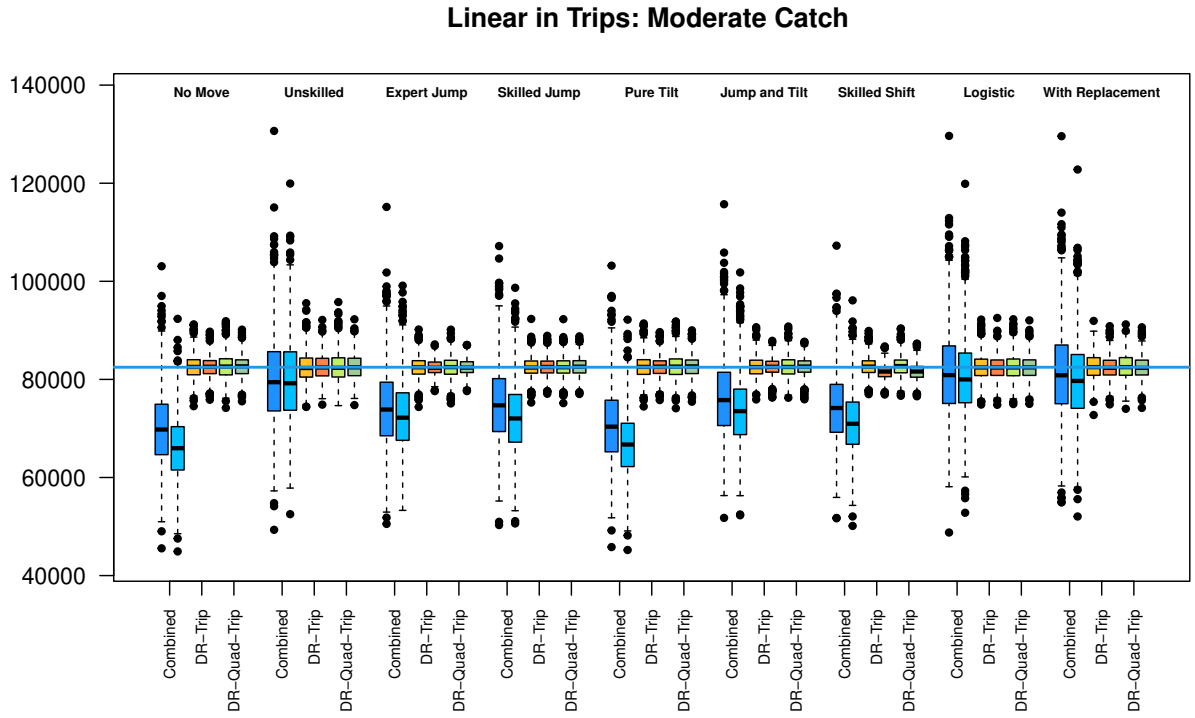

Figure 4.1: Boxplots of Linear in Trips: Moderate Catch estimates under 54 strategies. Each pair of successive boxplots corresponds to point estimates for one judgment behavior, and one estimator type under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot). Combined-Po is the combined estimator with pseudo log-likelihood assuming Poisson sampling; DR-Trip is the doubly-robust estimator with model linear in trips; DR-Quad-Trip is the doubly-robust estimator with model quadratic in trips.

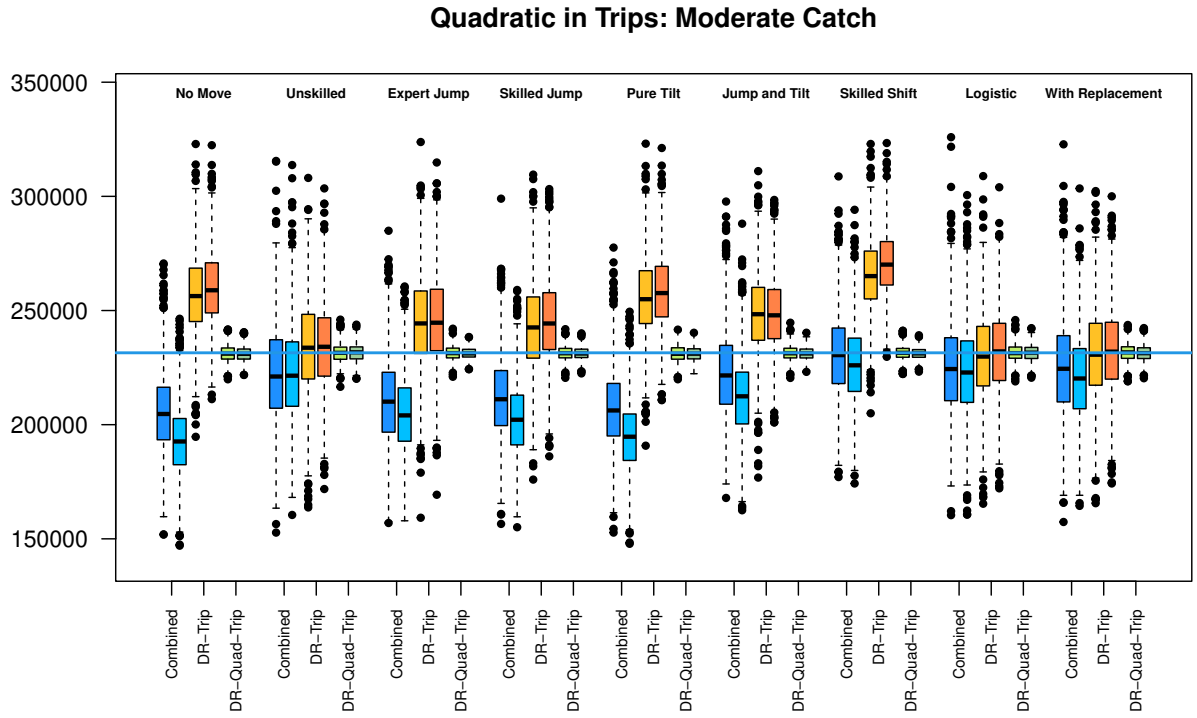

Figure 4.2: Boxplots of Quadratic in Trips: Moderate Catch estimates under 54 strategies. Each pair of successive boxplots corresponds to point estimates for one judgment behavior, and one estimator type under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot). Combined-Po is the combined estimator with pseudo log-likelihood assuming Poisson sampling; DR-Trip is the doubly-robust estimator with model linear in trips; DR-Quad-Trip is the doubly-robust estimator with model quadratic in trips.

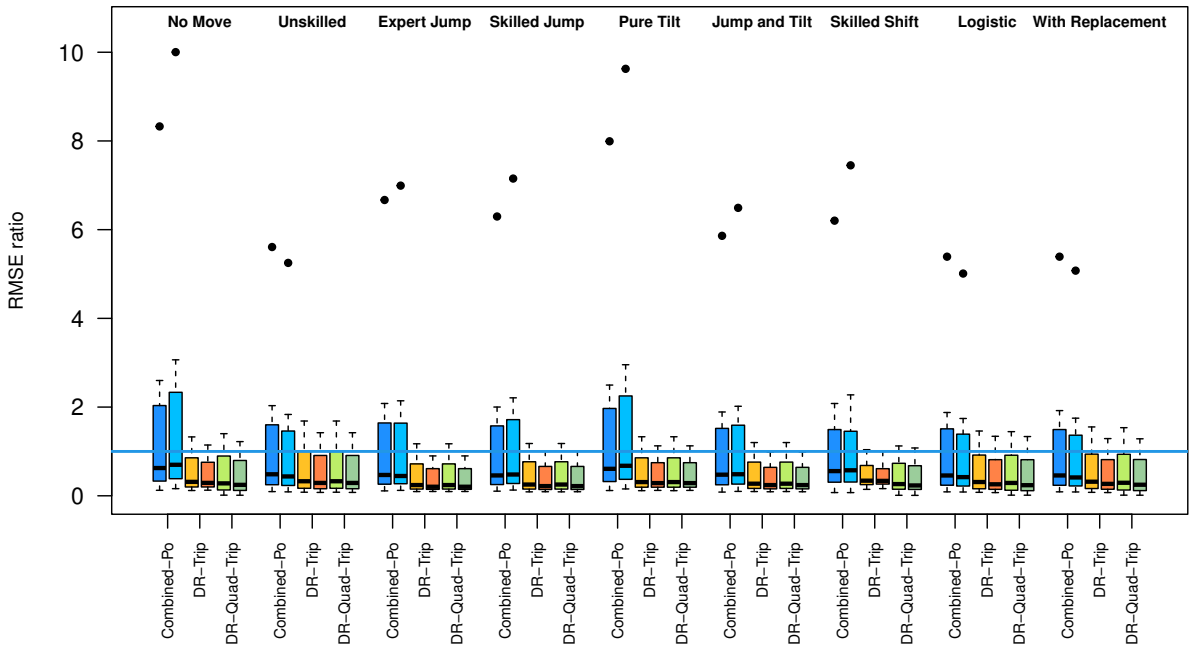

Figure 4.3: Ratio of RMSE for each strategy to RMSE of baseline strategy for catch across 54 strategies and 11 species. Values greater than one favor the baseline strategy, which is the no-move behavior regression estimator with original weights. Each pair of successive boxplots corresponds to RMSE ratios for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot). Combined-Po is the combined estimator with pseudo log-likelihood assuming Poisson sampling; DR-Trip is the doubly-robust estimator with model linear in trips; DR-Quad-Trip is the doubly-robust estimator with model quadratic in trips.

has smaller RMSE and lower bias compared to the combined estimator. Doubly-robust estimator fixes some bias from the misspecified inclusion probability model.

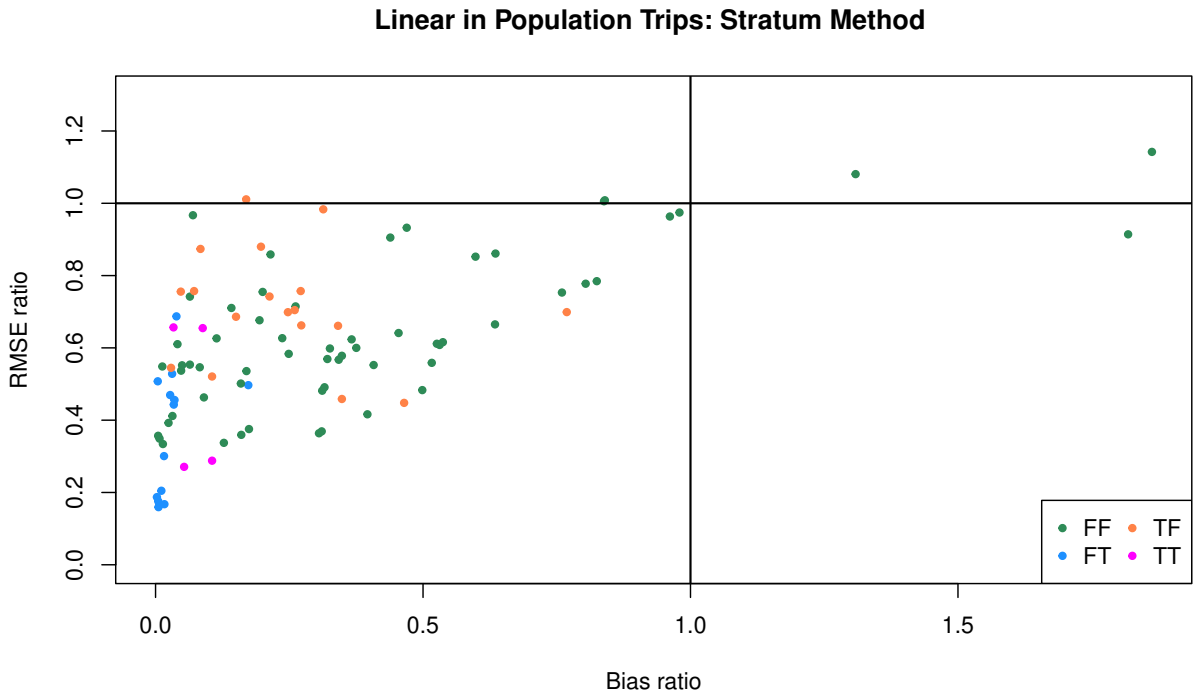

Figure 4.4: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch for the model linear in trips and stratum movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

We also summarize the results of catch rate similar to catch. Figure 4.8 summarizes all results as boxplots of the root mean square error (RMSE) ratios for estimated catch rate for the 11 species. In the ratios, combined-Po, DR-Trip, DR-Quad-Trip is the numerator and RMSE for no-move with original sample weights of regression estimator is the denominator. There are two boxplots for each judgment behavior and each estimator, with the left boxplot corresponding to the stratum movement method and the right boxplot corresponding to the bucket movement method. Figure 4.8

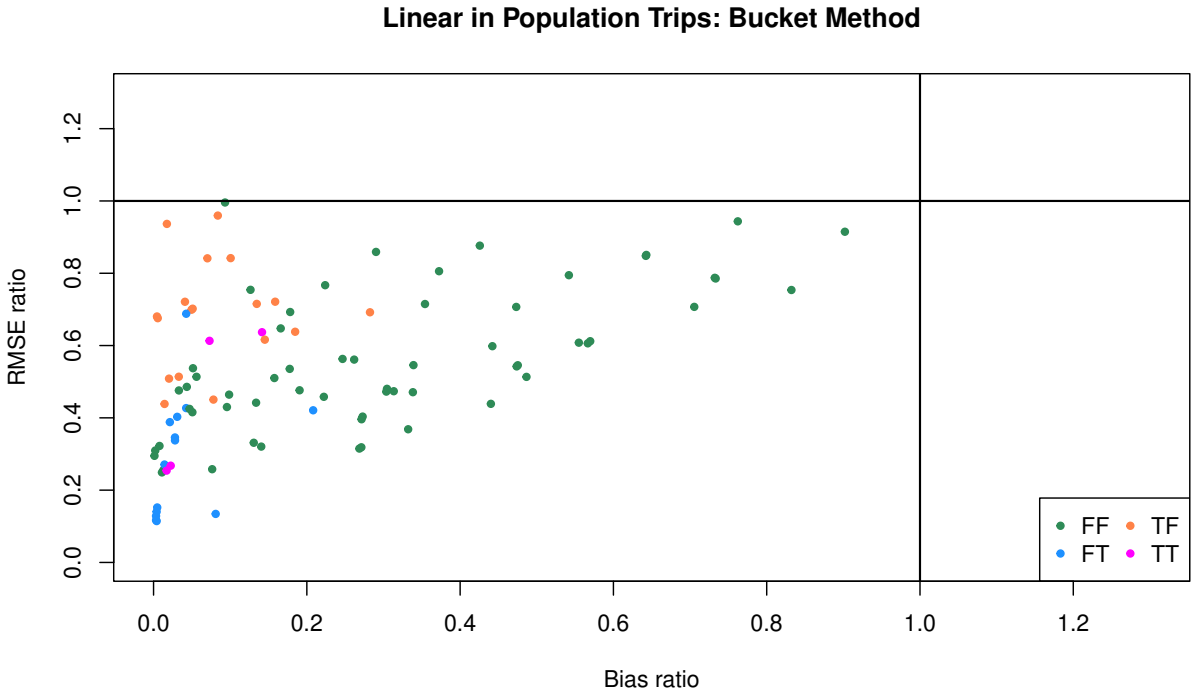

Figure 4.5: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch for the model linear in trips and bucket movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

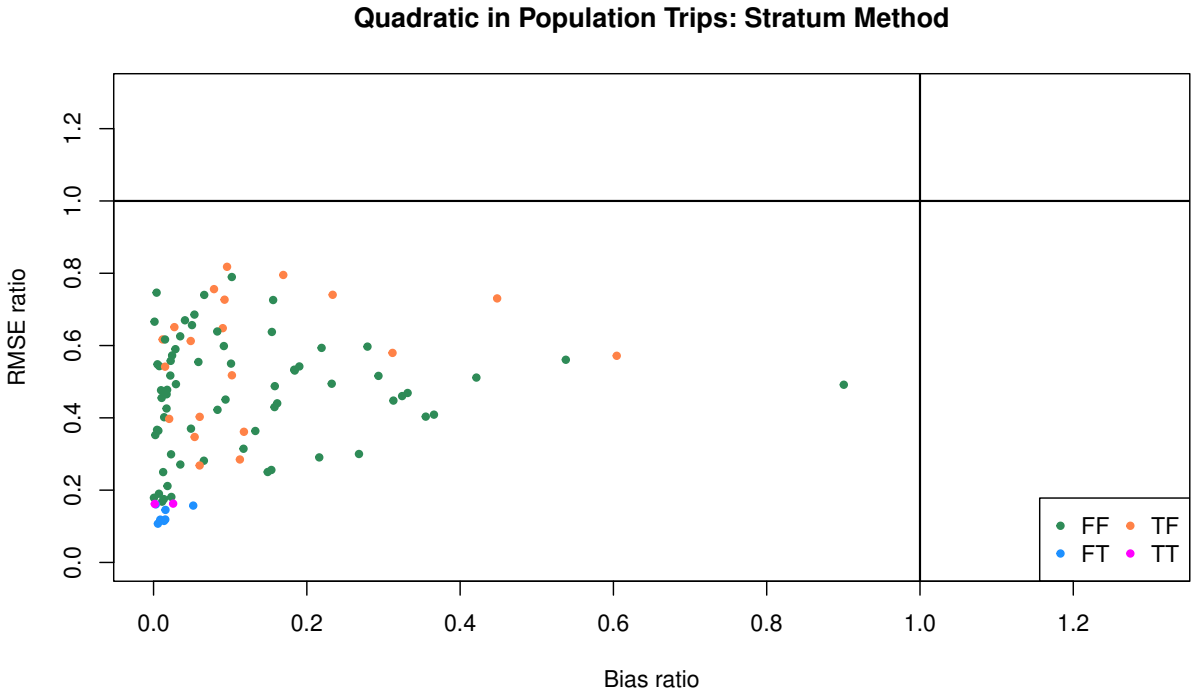

Figure 4.6: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch for the model quadratic in trips and stratum movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

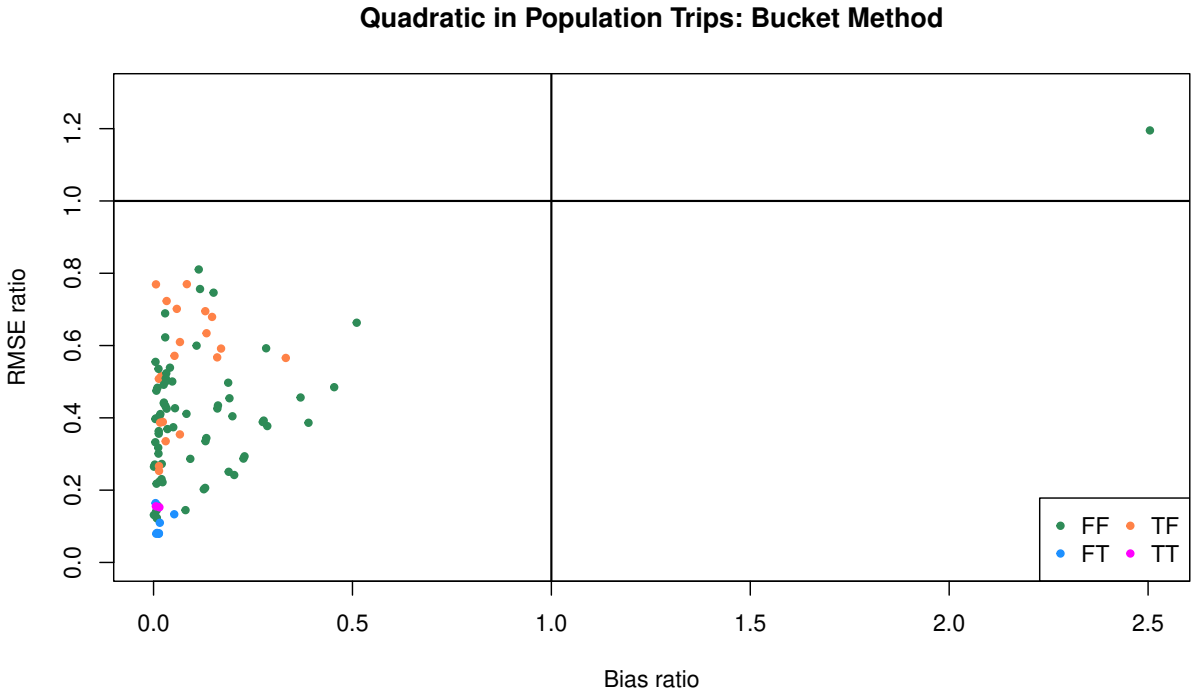

Figure 4.7: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch for the model quadratic in trips and bucket movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

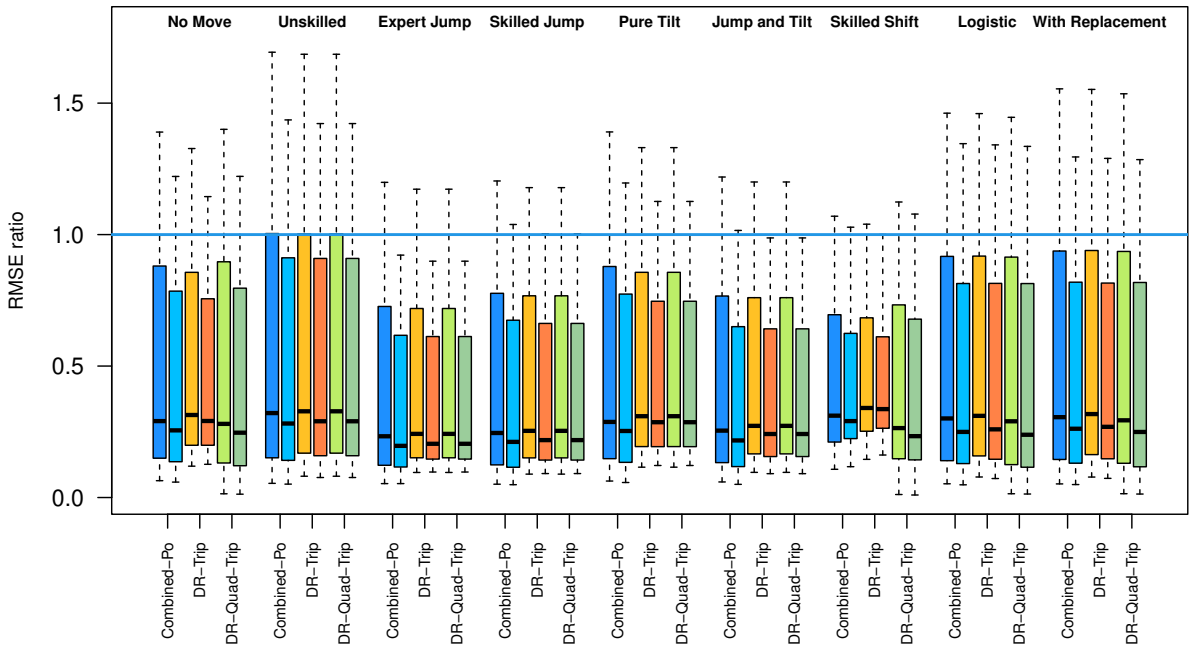

Figure 4.8: Ratio of RMSE for each strategy to RMSE of baseline strategy for catch rate across 54 strategies and 11 species. Values greater than one favor the baseline strategy, which uses the no-move behavior regression estimator with the original sample weights. Each pair of successive boxplots corresponds to RMSE ratios for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot). Combined-Po is the combined estimator with pseudo log-likelihood assuming Poisson sampling; DR-Trip is the doubly-robust estimator with model linear in trips; DR-Trip is the doubly-robust estimator with model quadratic in trips.

shows that the combined and doubly-robust estimator of catch rate also dominate the regression estimator across nearly all judgment behaviors and species. Figure 4.9 to Figure 4.12 summarize the results of the bias ratio and root mean square error (RMSE) ratios for the estimated catch rate for all 11 species. Figure 4.9 and Figure 4.10 are the summary of ratio with prediction model linear in trips. Figure 4.11 and Figure 4.12 are the summary of ratio with prediction model quadratic in trips. In the ratios, the absolute bias or RMSE for the doubly-robust estimator is the numerator and absolute bias or RMSE for the combined estimator is the denominator. Since the catch rate has doubly-robustness property, the doubly-robust estimator would not gain much by adding the prediction model. Figure 4.9 and Figure 4.10 show that we do not reduce the bias of catch rate from the doubly-robust estimator of the correctly specified regression model. Figure 4.11 and Fig-

ure 4.12 show that the doubly-robust estimator dominates the combined estimator if the regression model has quadratic term in trips and correctly specified.

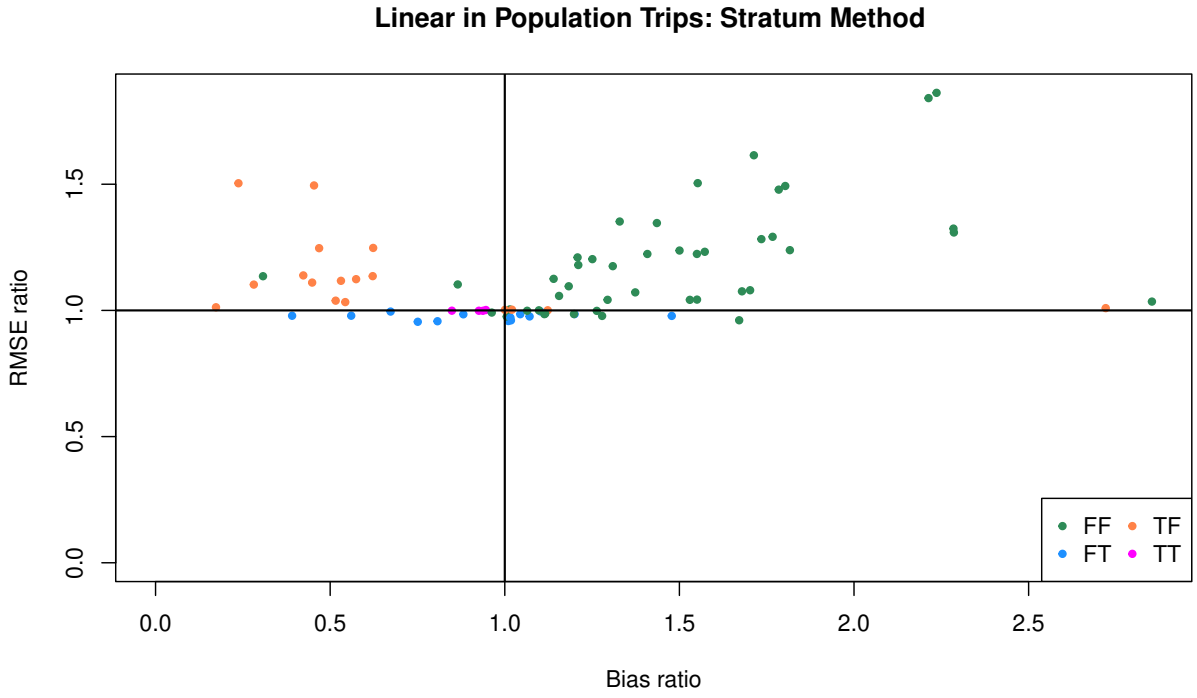

Figure 4.9: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch rate for the model linear in trips and stratum movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

We consider three different variance estimation methods discussed in Section 4.1.2 for the doubly-robust estimator of catch and catch rate with prediction model quadratic in trips since the the doubly-robust estimator works better than combined estimator for catch and catch rate that has quadratic term of trips. Figure 4.13 and 4.14 summarize the results as boxplots of the relative root mean square error (RMSE) for variance estimate for the 11 species of catch estimates. In the ratios, RMSE of the estimated standard deviation is the numerator and the true standard deviation

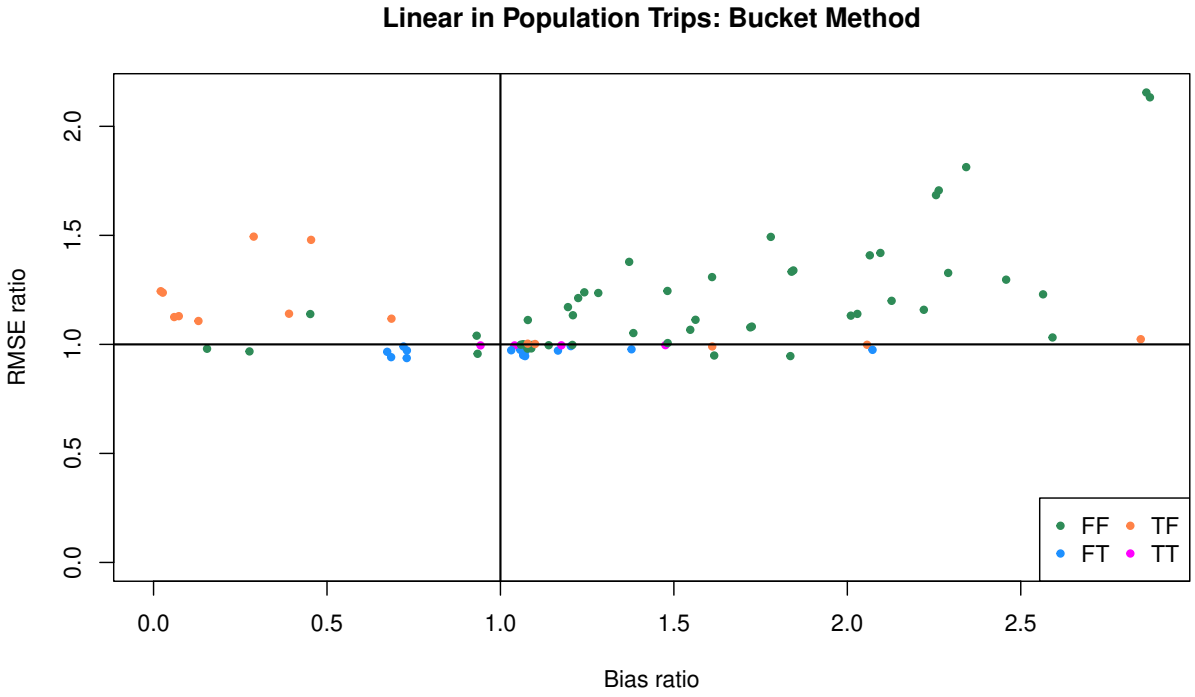

Figure 4.10: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch rate for the model linear in trips and bucket movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

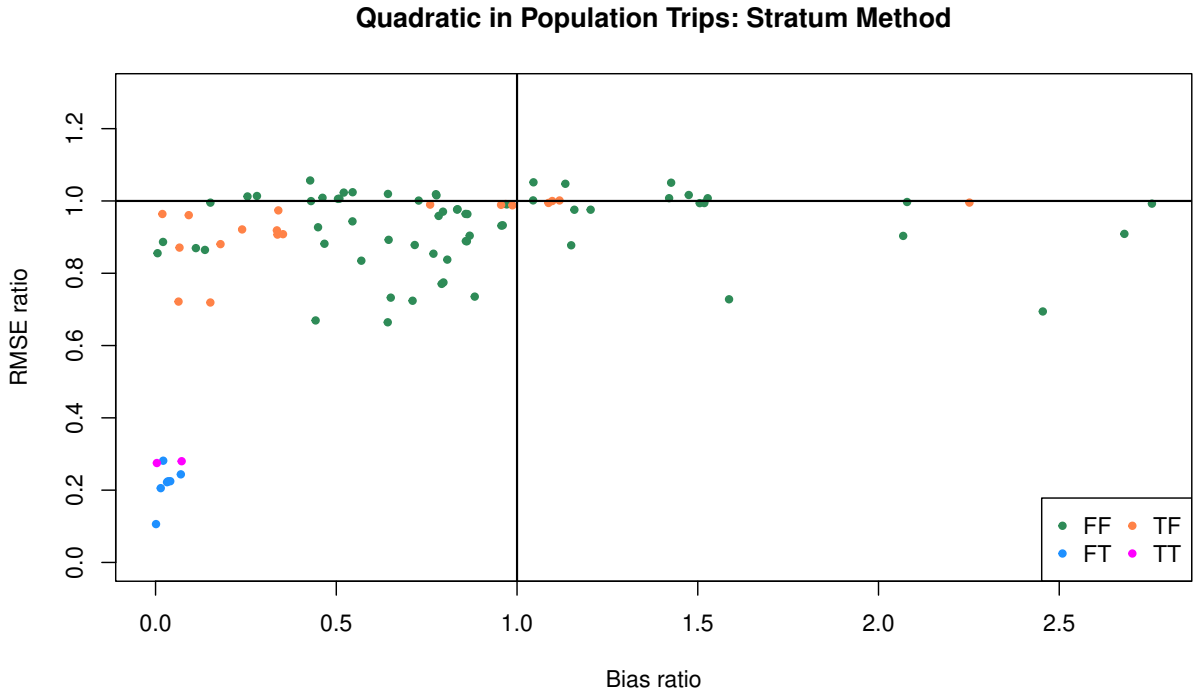

Figure 4.11: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch rate for the model quadratic in trips and stratum movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

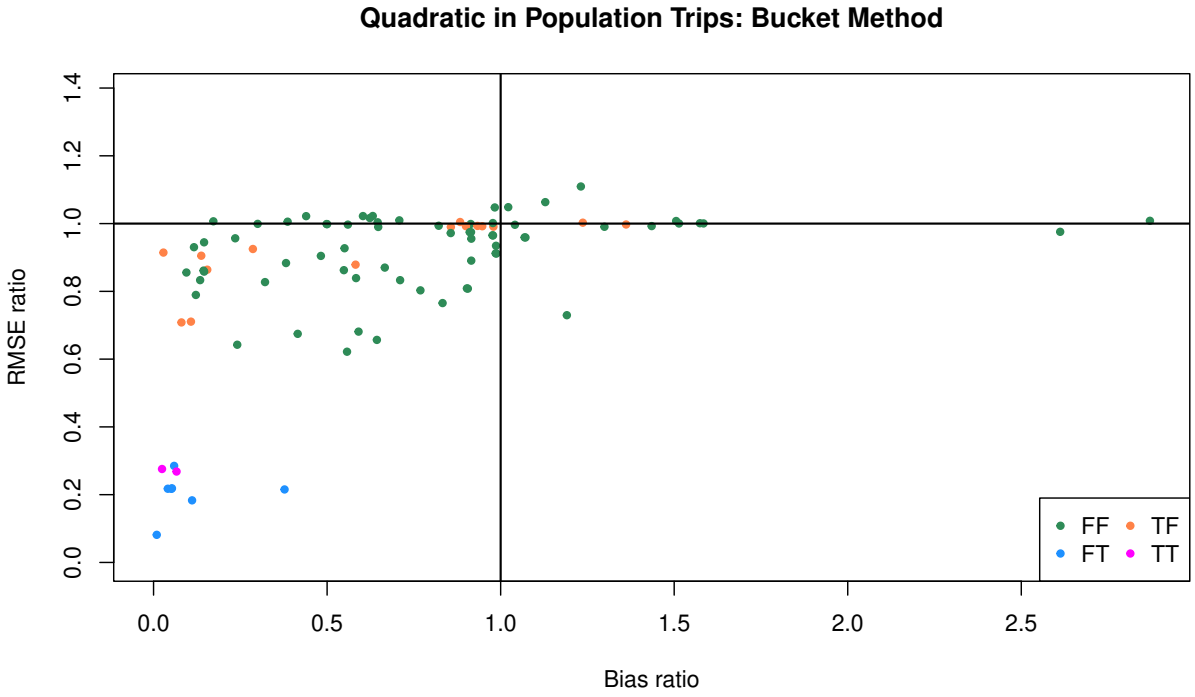

Figure 4.12: RMSE ratio versus bias ratio across 11 catch types and 9 judgment behaviors of catch rate for the model quadratic in trips and bucket movement method. Horizontal axis: ratio of doubly-robust estimator bias to combined estimator bias. Vertical axis: ratio of doubly-robust estimator RMSE to combined estimator RMSE. Values smaller than one favor the doubly-robust estimator. FF: cases for which both the probability and prediction models are misspecified; FT: cases for which the probability model is misspecified and the prediction model is correctly specified; TF: cases for which the probability model is correctly specified and the prediction model is misspecified; TT: cases for which both the probability and prediction model are correctly specified.

approximated via Monte Carlo is the denominator. There are two boxplots for each judgment behavior and each estimator, with the left boxplot corresponding to the stratum movement method and the right boxplot corresponding to the bucket movement method. The variance estimator $\widehat{V}_1$ is recommended since it is easily implemented in standard software and closer to the truth across the range of judgment behaviors, catch characteristics, and two moving methods. Figure 4.15 and 4.16 summarize the coverage of the 95% confidence interval for the 11 species of catch except for skilled shift judgment behavior, which is highly under-coverage due to the biased point estimate. The variance estimator $\widehat{V}_1$ also has better coverage across different catch characteristics and two moving methods. Figure 4.17 and 4.18 are the relative root mean square error (RMSE) for variance estimate for the 11 species of catch rate and 95% confidence interval coverage of catch rate using the quadratic in trips prediction model. These two figures also suggest $\widehat{V}_1$ has better coverage across different catch characteristics and judgment behavior.

4.2 A dual-frame approach for estimation of respondent-driven samples

4.2.1 Introduction

There are several possible applications of the estimation constructed in chapter 3. We consider applying the methodology to the respondent-driven samples. Respondent Driven Sampling (RDS) is a chain-referral sampling that is implemented to study the hidden population and introduced by Heckathorn (1997). RDS sampling starts with seeds from the target population and progresses with recruiting waves until the desired sample size is reached. If seeds are selected according to some probability design, the seeds can be treated as strict probability samples, and those being recruited are nonprobability samples since we do not know the probability of the recruitment process. In the existing literature, several assumptions are needed to get an asymptotically unbiased estimate but these assumptions might be unrealistic. In this section, we apply the dual-frame estimator in chapter 3 to the RDS sample.

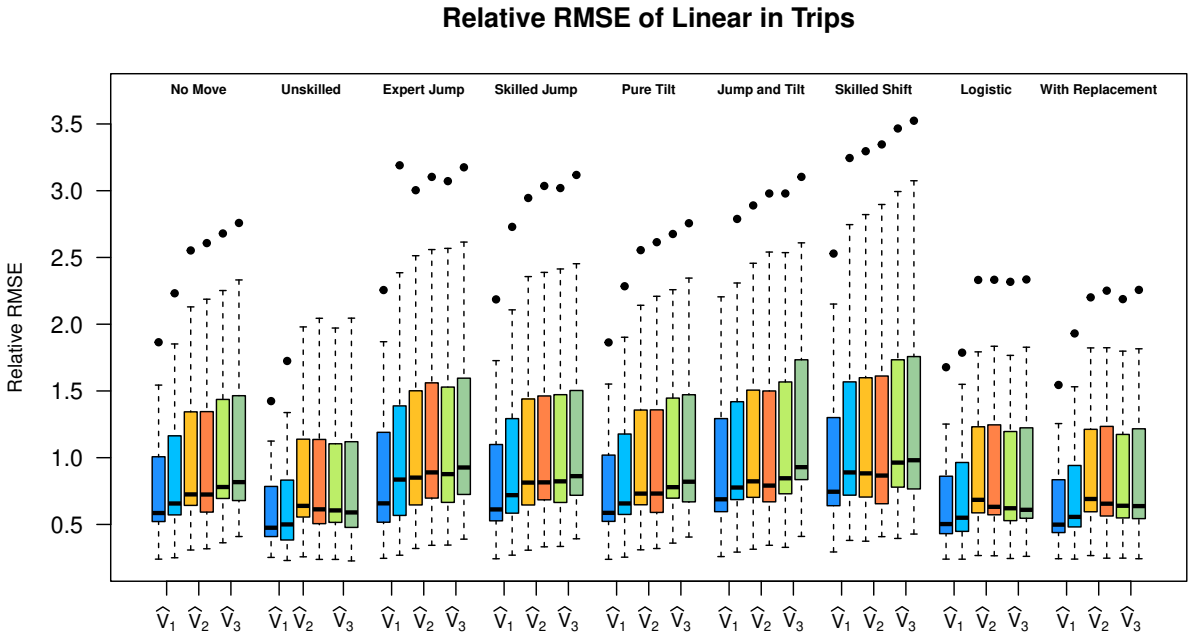

Figure 4.13: Relative RMSE of standard deviation for catch using three different variance estimator across 54 strategies and 11 species for the doubly-robust estimator with model linear in trips. Smaller values are better. Each pair of successive boxplots corresponds to relative RMSE for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

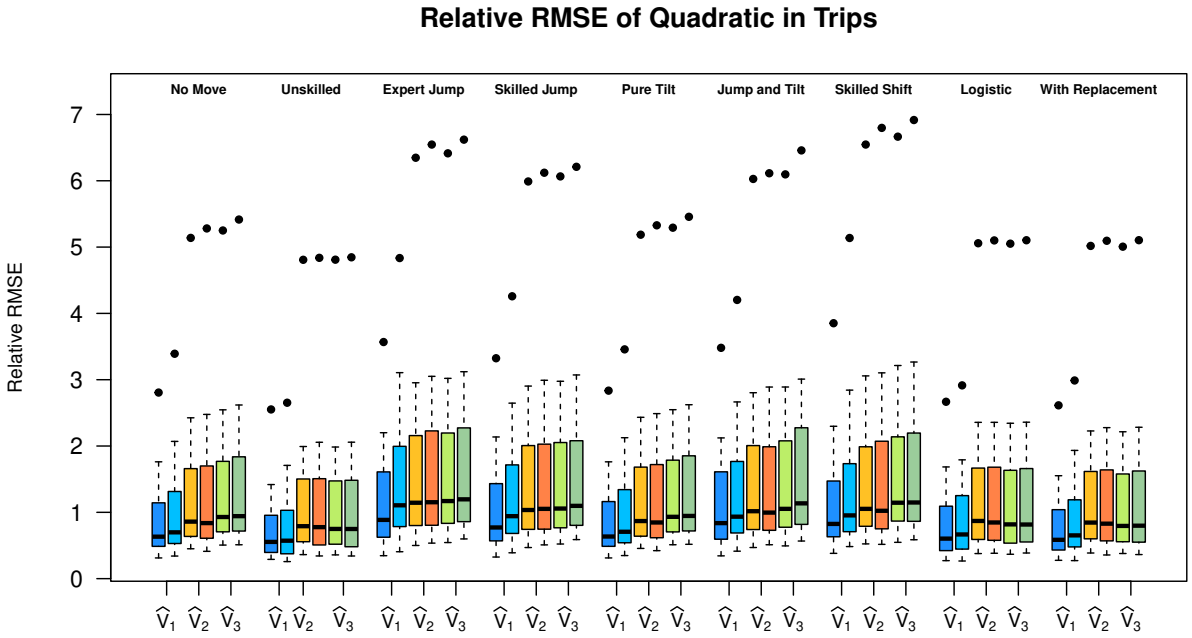

Figure 4.14: Relative RMSE of standard deviation for catch using three different variance estimator across 54 strategies and 11 species for the doubly-robust estimator with model quadratic in trips. Smaller values are better. Each pair of successive boxplots corresponds to relative RMSE for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

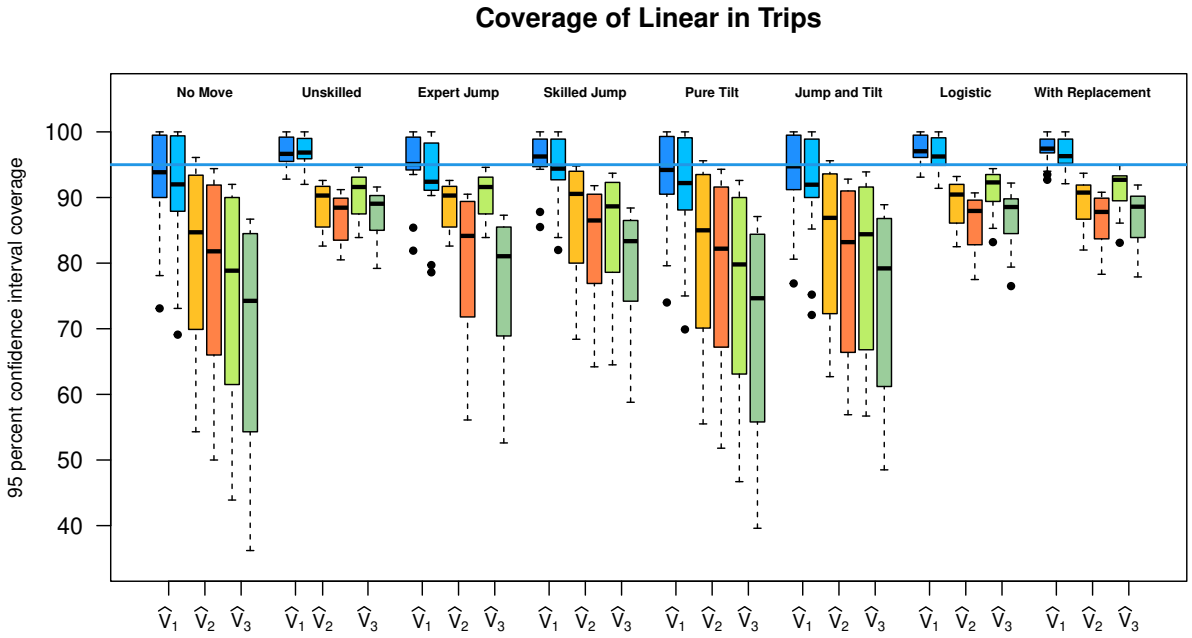

Figure 4.15: 95 percent confidence interval coverage for catch across 48 strategies and 11 species of each variance estimate with model linear in trips. Each pair of successive boxplots corresponds to confidence interval coverage for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

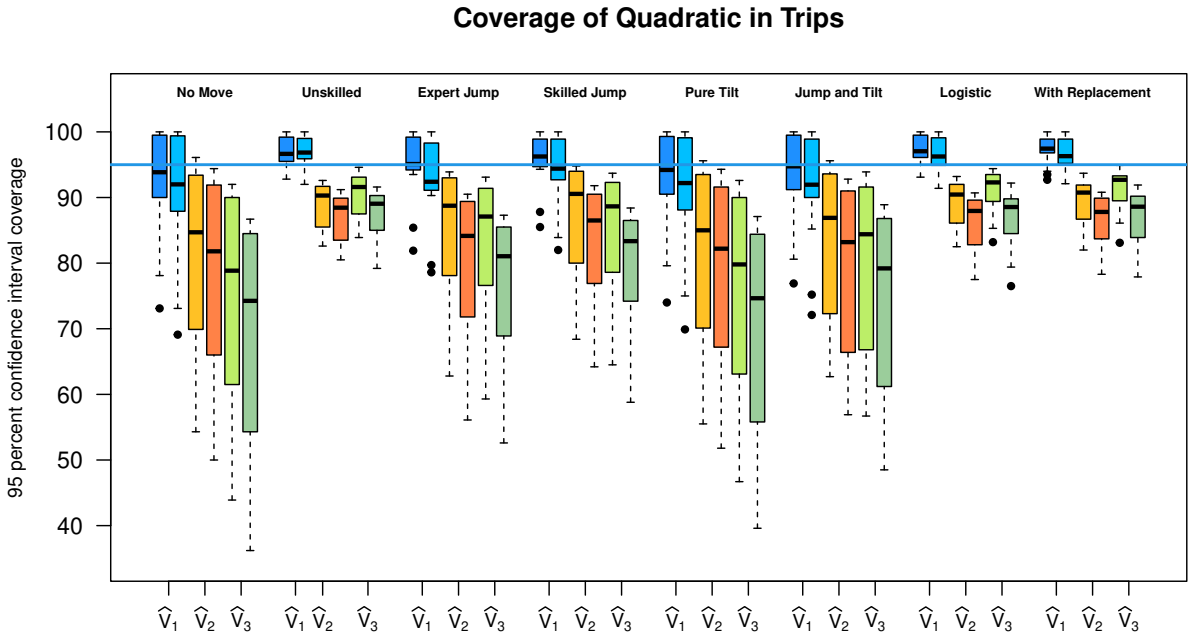

Figure 4.16: 95 percent confidence interval coverage for catch across 48 strategies and 11 species of each variance estimate with model quadratic in trips. Each pair of successive boxplots corresponds to confidence interval coverage for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

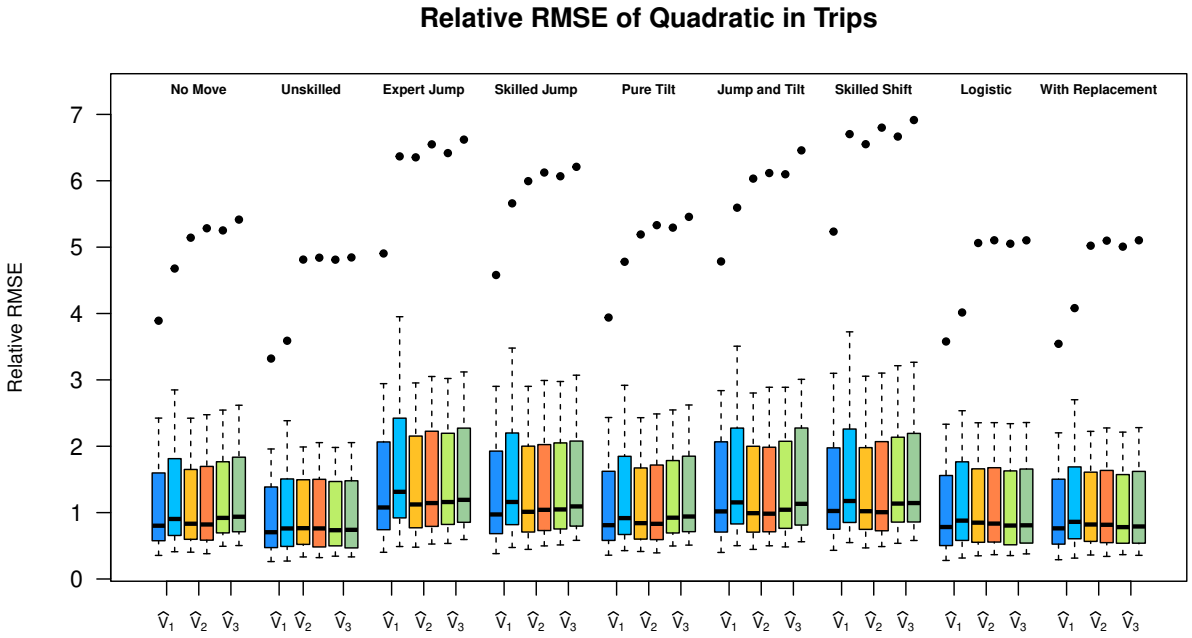

Figure 4.17: Relative RMSE of standard deviation for catch rate using three different variance estimator across 54 strategies and 11 species for the doubly-robust estimator with model quadratic in trips. Smaller values are better. Each pair of successive boxplots corresponds to relative RMSE for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

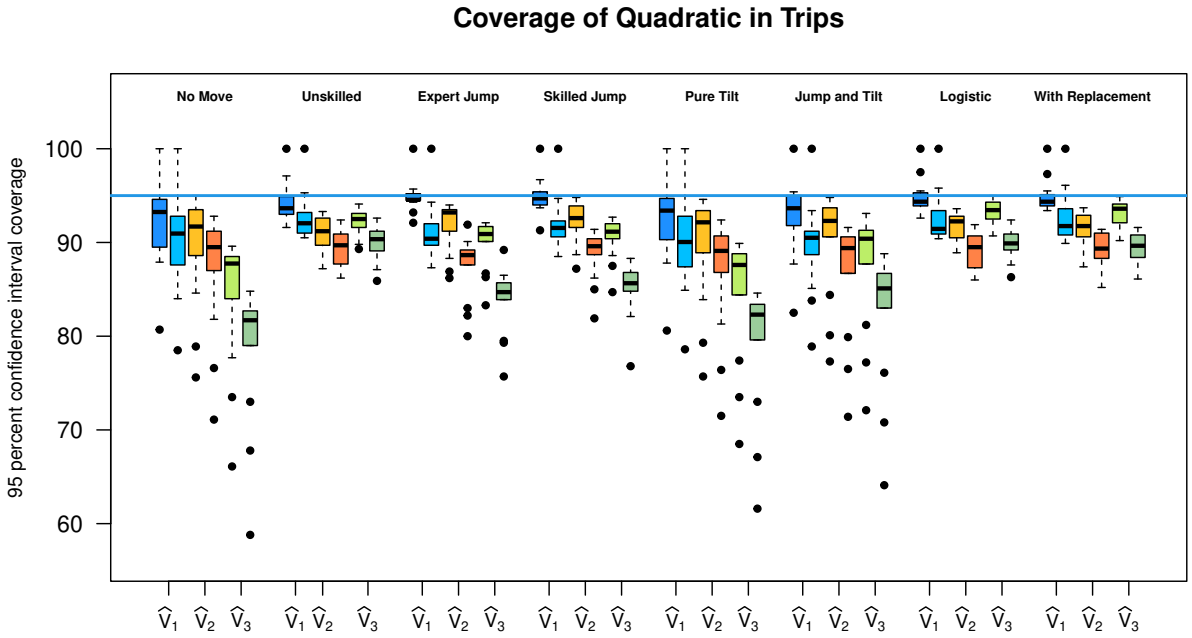

Figure 4.18: 95 percent confidence interval coverage for catch rate across 48 strategies and 11 species of each variance estimate with model quadratic in trips. Each pair of successive boxplots corresponds to confidence interval coverage for one judgment behavior, one estimator type, and 11 species under the stratum movement method (left boxplot) and under the bucket movement method (right boxplot).

4.2.2 Dual-frame methods applied to the respondent-driven sampling

In order to construct a scenario similar to the judgment sample, we consider the case in which the seeds are selected by some probability design with known inclusion probabilities. This is also the case considered in Michaels et al. (2019). The individuals are sampled without replacement. The only assumption we make is Poisson sampling in estimating the inclusion probability for the nonprobability samples. We make no assumptions about the population network. Let s_A denote the seeds and s_B denote the individuals being recruited. We consider without-replacement sampling $s_A \cap s_B = \emptyset$, $s = s_A \cup s_B$. Let n_A denote the number of seeds, and n_B is the number of individuals being recruited. We apply the same likelihood and normalization process in section 3.2 to estimate the inclusion probability for the individuals being recruited. The combined estimators for the total and the mean of the variable of interest are

$$\widehat{T}_{y,\text{com}} = \sum_{k \in s} \frac{y_k}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k},$$

$$\widehat{\mu}_{y,\text{com}} = \sum_{k \in s} \frac{y_k/\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k}{1/\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k}, \tag{4.7}$$

where $\widehat{\rho}_k$ is the estimated inclusion probability for the individual being recruited.

The most commonly used estimator in the current RDS is the VH estimator, and Gile and Handcock (2010) state that the VH estimator performs better than the SH estimator. The combined estimator would be less efficient if the nonprobability sample size is large. For these reasons, we also construct the convex combination of the VH estimator and the combined estimator of the total and mean

$$\begin{aligned}
\widehat{T}_{y,\text{convex}} &= \frac{n_A}{n_A+n_B} \sum_{k \in s} \frac{y_k}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k} + \frac{n_B}{n_A+n_B} \sum_{k \in s} \frac{N d_k^{-1} y_k}{\sum_{k \in s} d_k^{-1}} \\
&= \sum_{k \in s} \left[\frac{n_A}{n_A+n_B} \frac{1}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k} + \frac{n_B}{n_A+n_B} \frac{N d_k^{-1}}{\sum_{k \in s} d_k^{-1}} \right] y_k,
\end{aligned}$$

$$\widehat{\mu}_{y,\text{convex}} = \frac{\sum_{k\in s}\left[n_A/\left\{(n_A+n_B)\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)\right\}+n_BNd_k^{-1}/\left\{(n_A+n_B)\left(\sum_{k\in s}d_k^{-1}\right)\right\}\right]y_k}{\sum_{k\in s}\left[n_A/\left\{(n_A+n_B)\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)\right\}+n_BNd_k^{-1}/\left\{(n_A+n_B)\left(\sum_{k\in s}d_k^{-1}\right)\right\}\right]}. \tag{4.8}$$

The variance and variance estimator of (4.7) can be approximated by Taylor expansion and $\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k$ can be treated as the corresponding combined inclusion probability. Similarly, the variance and variance estimator of (4.8) can be approximated by Taylor expansion and

$$n_A/\left\{(n_A+n_B)\left(\pi_k^A+(1-\pi_k^A)\widehat{\rho}_k\right)\right\}+n_BNd_k^{-1}/\left\{(n_A+n_B)\left(\sum_{k\in s}d_k^{-1}\right)\right\}$$

can be treated as the corresponding combined inclusion probability.

4.2.3 Simulation experiment

The proposed estimators are evaluated and compared to the three current estimators SH, VH, and SS using the Project 90 study. The data was collected between 1988 and 1992 in Colorado Springs, CO to study the heterosexuals' transmission of HIV, and is network data on the hidden population. Several published studies (Baraff et al., 2016; Fellows, 2019; Goel and Salganik, 2010) have used Project 90 data to compare the RDS estimators. As in the prior studies, we constructed a subset of the network containing the largest connected component, which includes 4430 individuals and 18407 edges. The data includes 13 individual attributes such as sex worker, pimp, drug dealer. The values of attributes are binary, and 1 indicates the individual has the attribute. Table 4.1 summarizes all the 13 attributes and the population proportion.

We consider two target sample sizes, 130 and 150, each with 100 seeds. These targets reflect the numbers from Michaels et al. (2019). The design is without replacement sampling and seeds are given three coupons. We consider two selections of seeds, proportional to degree or randomly, and eight different recruitment behaviors similar to having different judgment behaviors: (1) *random*, in which three acquaintances are recruited at random with equal probabilities, if possible; (2) *recruit fraction*, in which 0, 1, 2, or 3 acquaintances are recruited at random, with probabilities $(1/6, 1/6, 1/6, 1/2)$; (3) *degree*, in which recruitment probabilities are proportional to the degrees

Table 4.1: Project 90 population proportion

Attributes	Population proportion
female	0.43
sex worker	0.06
pimp	0.02
client	0.10
drug dealer	0.08
drug cook	0.01
thief	0.03
retired	0.03
housewife	0.06
disabled	0.04
unemployed	0.17
homeless	0.14
nonwhite	0.26

of acquaintances; (4) *inverse degree*, in which recruitment probabilities are proportional to the inverse degrees of acquaintances; (5) *prefer female*, in which females must recruit female acquaintances, if possible, and males recruit males; (6) *prefer pimp*, in which pimps must recruit pimp acquaintances, if possible, and non-pimps recruit non-pimps; (7) *expert female*, in which everyone must recruit female acquaintances, if possible; and (8) *expert pimp*, in which everyone must recruit pimp acquaintances, if possible. In the existing literature, recruitment is assumed at random, but our estimator allows for differential recruitment.

For all the recruitment behaviors, we estimate the inclusion probabilities using the model

$$\text{logit}(\rho_k) = \beta_0 + \beta_1 \text{degree}$$

so the probabilities are the same across all attributes. The model is misspecified for all the simulated recruitment behaviors, though it is somewhat similar to (3) degree. We draw 1000 samples, use all recruitment behaviors, estimate the inclusion probabilities assuming Poisson sampling, construct five different estimators (SH, VH, SS, combined, and convex combination of VH and combined), and apply these estimators to thirteen attributes. The estimates, standard deviations, and

the confidence intervals of SH, VH, and SS estimator are computed using the `R` package "RDS" (Handcock et al., 2021).

Figure 4.19 and Figure 4.20 summarize all results as boxplots of the root mean square error (RMSE) ratios for 13 attributes. In the ratios, RMSE for combined estimator is the numerator and RMSE for SH, VH, SS, convex combination is the denominator. There are two boxplots in each estimator and recruitment behavior, the left one is 130 sample size, and the right one is 150 sample size. Both figures show that combined estimator or convex combination reduces the mean square error. To help see the improvement of RMSE in the combined estimator, we also rank the five estimators and summarize the ranking across all attributes and recruitment behaviors in Table 4.2. The combined estimator has the lowest average rank for different sample sizes. Combined estimator or convex combination is doing better than the existing estimators.

Table 4.2: Summary results for average rank across all attributes of five different estimators, SH, VH, SS, Combined, and Convex combination for two different sample size with two kinds of seeds selection.

		Estimator				
		SH	VH	SS	Combined	Convex
Seeds proportional to degree	130 sample size	5	3.9	3.02	1.19	1.88
	150 sample size	5	3.82	3.03	1.27	1.88
Seeds randomly	130 sample size	5	3.94	2.94	1.73	1.38
	150 sample size	5	3.94	2.93	1.55	1.58

Figure 4.21 and Figure 4.22 summarize the results as boxplots of the relative root mean square error (RMSE) for variance estimate for different recruitment behaviors across 13 attributes. In the ratios, RMSE of the estimated standard deviation is the numerator and the true standard deviation approximated via Monte Carlo is the denominator. Figure 4.23 and Figure 4.24 summarize the coverage of the 95% confidence interval for different recruitment behaviors across 13 attributes. Combined estimator and convex combination are closer to the true standard deviation approximated via Monte Carlo and nominal 95 percent confidence interval coverage.

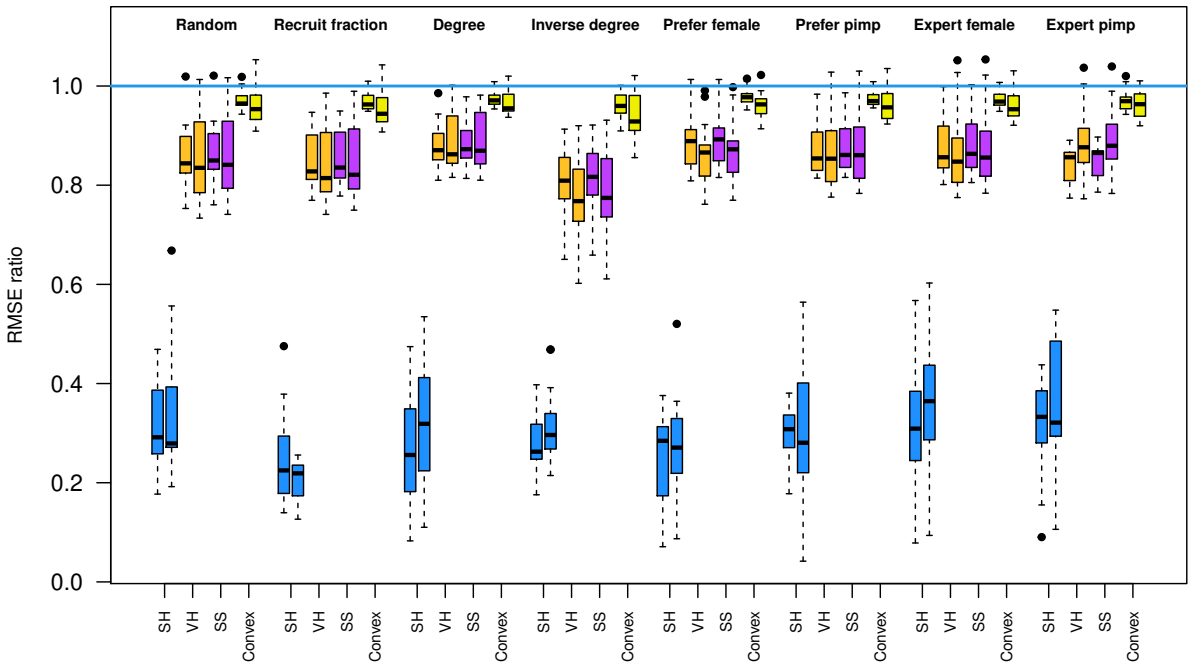

Figure 4.19: Ratio of RMSE for each recruitment behavior to RMSE of combined estimator across 13 attributes of seeds selected proportional to degree. Values smaller than one favor the combined estimator. Each pair of successive boxplots corresponds to RMSE ratios for one recruitment behavior, one estimator type, and 13 attributes under 130 sample size (left boxplot) and 150 sample size (right boxplot).

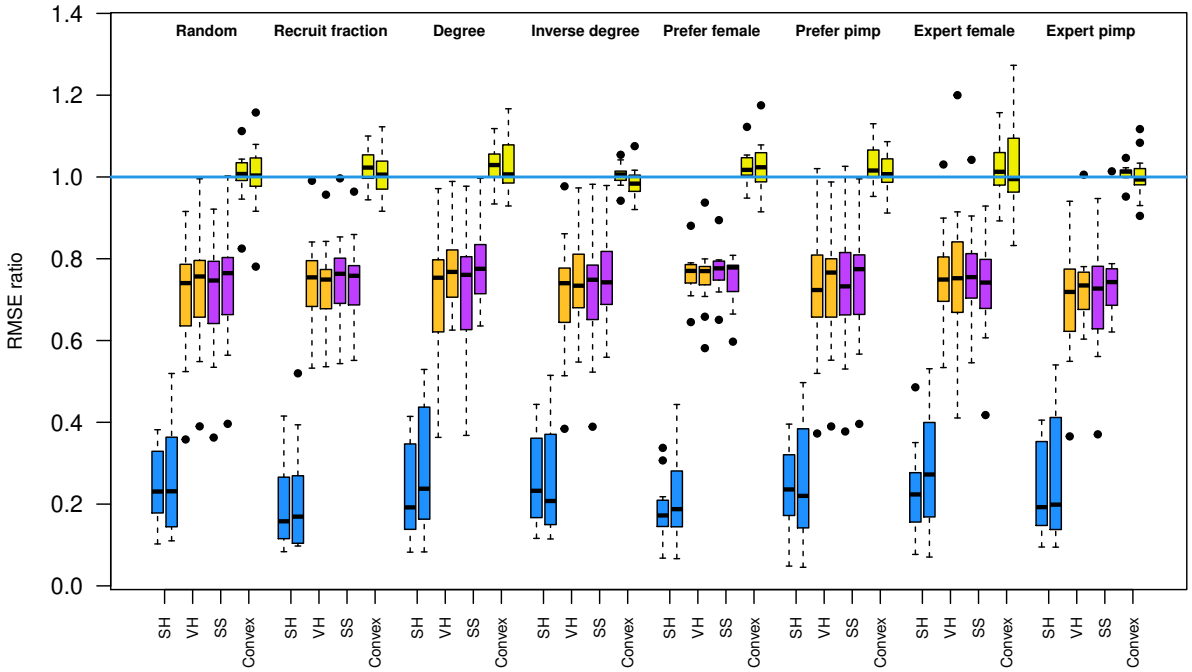

Figure 4.20: Ratio of RMSE for each recruitment behavior to RMSE of combined estimator across 13 attributes of seeds selected randomly. Values smaller than one favor the combined estimator. Each pair of successive boxplots corresponds to RMSE ratios for one recruitment behavior, one estimator type, and 13 attributes under 130 sample size (left boxplot) and 150 sample size (right boxplot).

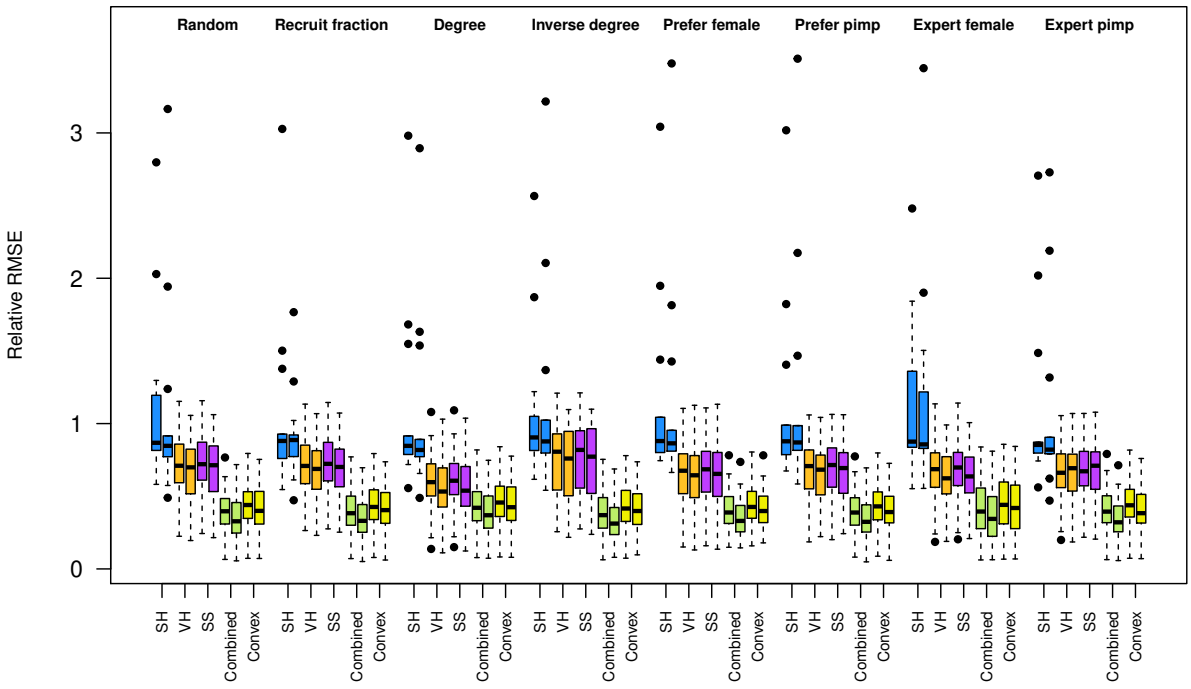

Figure 4.21: Relative RMSE of standard deviation of each strategy across 13 attributes for five estimator of seeds selected proportional to degree. Smaller values are better. Each pair of successive boxplots corresponds to relative RMSE for one recruitment behavior, one estimator type, and 13 attributes under 130 sample size (left boxplot) and 150 sample size (right boxplot).

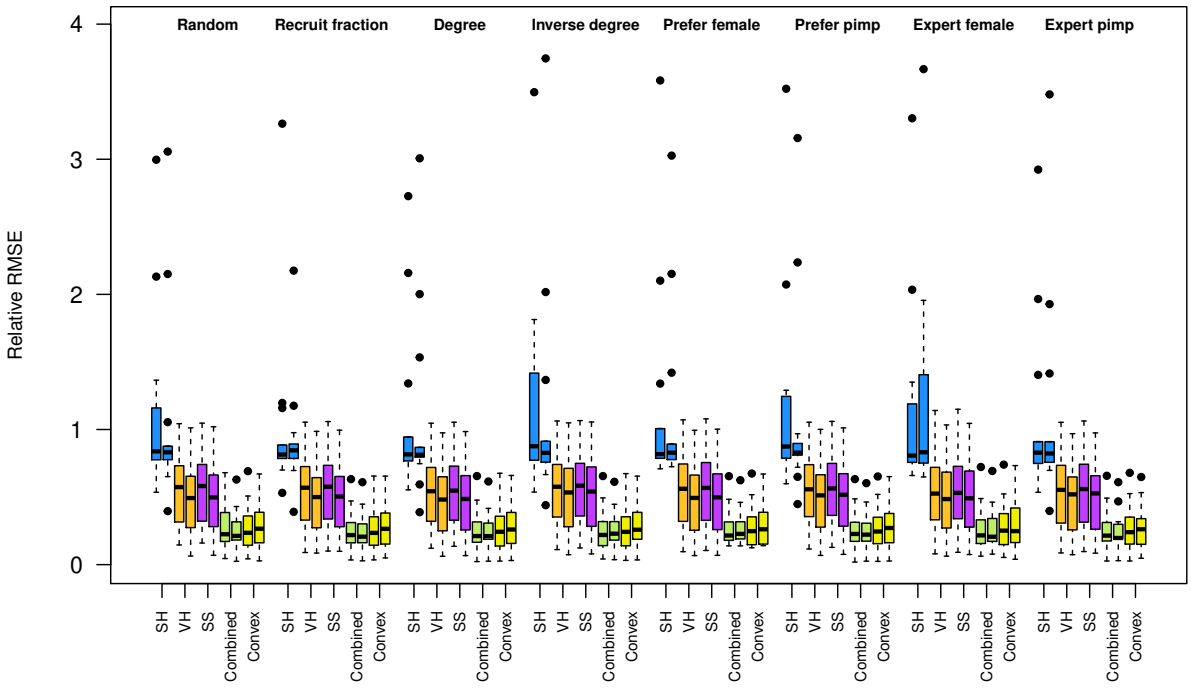

Figure 4.22: Relative RMSE of standard deviation of each strategy across 13 attributes for five estimator of seeds selected randomly. Smaller values are better. Each pair of successive boxplots corresponds to relative RMSE for one recruitment behavior, one estimator type, and 13 attributes under 130 sample size (left boxplot) and 150 sample size (right boxplot).

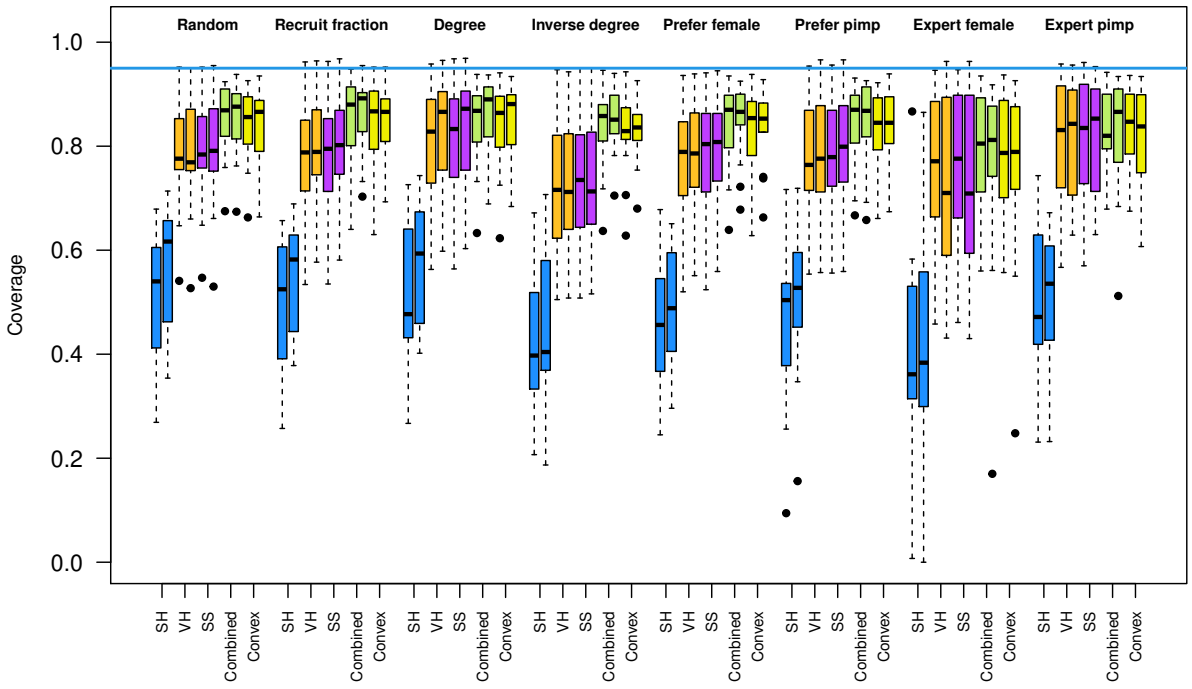

Figure 4.23: 95 percent confidence interval coverage of each strategy across 13 attributes for five estimator of seeds selected proportional to degree. Each pair of successive boxplots corresponds to confidence interval coverage for one recruitment behavior, one estimator type, and 13 attributes under 130 sample size (left boxplot) and 150 sample size (right boxplot).

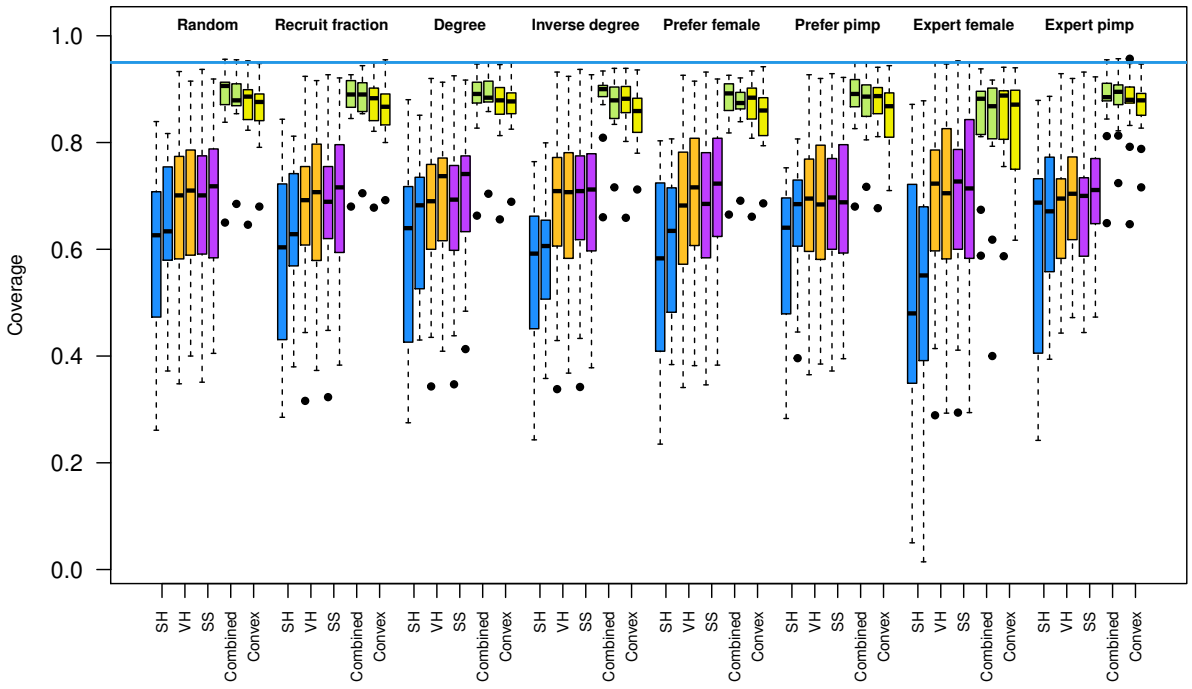

Figure 4.24: 95 percent confidence interval coverage of each strategy across 13 attributes for five estimator of seeds selected randomly. Each pair of successive boxplots corresponds to confidence interval coverage for one recruitment behavior, one estimator type, and 13 attributes under 130 sample size (left boxplot) and 150 sample size (right boxplot).

4.3 A dual-frame approach for combing probability and non-probability samples

4.3.1 Introduction

In chapter 3, we assume probability sample and nonprobability sample are selected from the same frame. However, this might not be true in many nonprobability sampling settings. In this section, we consider the extension that nonprobability samples are drawn from a subset of the whole population, so that the nonprobability samples have undercoverage bias. We utilize the frames and simulated samples described in Benoit-Bryan and Mulrow (2021), which were created to evaluate different estimation approaches that combine probability and nonprobability samples. The data used to generate these simulated samples is from a real study called Culture and Community in a Time of Crisis (CCTC), which evaluated behaviors and attitudes during the pandemic crisis. In the simulated version, the probability sample frame is the full population, and the nonprobability sample frame is from an unknown subset of the probability sample frame and has unknown selection mechanism. The probability sample is representative and small, the nonprobability sample is biased but large. Our goal is to estimate characteristics of the population of interest by combining the probability and nonprobability samples. The approach we used is a dual-frame technique adapted from chapter 3.

4.3.2 Estimation

We modified the combined estimator and separate estimator, introduced in chapter 3, to the scenario that the nonprobability sample frame is a subset of the probability sample frame. Let s_A denote a probability sample from U with known inclusion probabilities, and s_B denote a nonprobability sample from sub-population U_B with unknown inclusion probabilities. Any elements selected in both s_A and s_B are screened out in the s_B sample, so that $s_A \cap s_B = \emptyset$. The probability sample indicators are $I_k^A = 1$ if k is from the probability sample, $I_k^A = 0$ otherwise; $I_k^B = 1$ if k is from the nonprobability sample, $I_k^B = 0$ otherwise. The first order inclusion probability for s_A is

$\pi_k^A = \mathrm{P}\left[I_k^A = 1\right]$, where $\pi_k^A > 0$ and known for all $k \in s_A$. The first order inclusion probability for s_B is $\pi_k^B = \rho_k(1 - \pi_k^A)$. Because of the unknown selection mechanism, both ρ_k and π_k^A are unknown for $k \in s_B$.

We use statistical matching from the nonprobability sample to the probability sample to figure out what part of the universe s_B represents, and to determine π_k^A for $k \in s_B$. Each nonprobability sample element is matched to one probability sample element using Gower's distance measure (Gower, 1971). We define the matching notation $m_{k\ell} = 1$ if $k \in s_A$ is matched to unit $\ell \in s_B$, $m_{k\ell} = 0$ otherwise. If $\sum_{\ell \in s_B} m_{k\ell} = m_{k+} > 0$, some element in s_A is matched in s_B, and $m_{k+} = 0$ otherwise. Let $b_k = 1$ if $m_{k+} > 0$, $b_k = 0$ if $m_{k+} = 0$. If the nonprobablity sample size n_B is large, b_k is almost not random, hence we treat b_k as nonrandom here. Using a similar argument to that in chapter 3 for estimating ρ_k, we construct the pseudo log-likelihood function using only the matched part of s_A and s_B, and assuming Poisson sampling for s_B:

$$\sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho_k}{1-\rho_k}\right) + \sum_{k \in U} \ln\left(1-\rho_k\right)\left(1-\pi_k^A\right)\frac{I_k^A}{\pi_k^A} b_k.$$

If ρ_k follows a logistic model $\text{logit}\left(\rho_k\right) = \boldsymbol{x}_k^\top \boldsymbol{\theta}$, the pseudo log-likelihood is

$$\sum_{k \in U \setminus s_A} I_k^B \boldsymbol{x}_k^\top \boldsymbol{\theta} - \sum_{k \in U} (1-\pi_k^A)\frac{I_k^A}{\pi_k^A} b_k \ln\left\{1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta}\right)\right\}.$$

We also normalize the estimated ρ_k so that the expected nonprobability sample size matches the actual sample size.

The matching step helps us figure out the undercoverage of the nonprobability sample frame and π_k^A of the nonprobability samples. The matched part of the s_A sample ($m_{k+} > 0$) is covered by both the probability sampling frame and the nonprobability sampling frame. For the matched sample, we utilize the combined or separate estimator to estimate the total. For the unmatched part of the s_A sample ($m_{k+} = 0$), we use only the probability sample to estimate the total since the probability sample is the only data source. The modified combined estimator for the mean

$\widehat{R}_{y,\mathrm{com,m}}$ is then

$$\frac{\sum_{k\in U} y_k b_k (I_k^A + (1-I_k^A)I_k^B)/(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k) + \sum_{k\in U} y_k\,(1-b_k)\,I_k^A/\pi_k^A}{\sum_{k\in U} b_k (I_k^A + (1-I_k^A)I_k^B)/(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k) + \sum_{k\in U} (1-b_k)\,I_k^A/\pi_k^A}. \tag{4.9}$$

The design variance of the modified combined estimator can be approximated by Taylor expansion, and derived from the iterated variance by conditioning on the s_A sample first. The Taylor expansion for the ratio of the combined estimator is approximately

$$\begin{aligned}
& \frac{\sum_{k\in U} y_k}{\sum_{k\in U} 1} + \frac{1}{\sum_{k\in U} 1}\left[\sum_{k\in U} \frac{y_k b_k (I_k^A + (1-I_k^A)I_k^B)}{(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k)} + \sum_{k\in U} y_k\,(1-b_k)\,\frac{I_k^A}{\pi_k^A} - \sum_{k\in U} y_k\right] \\
+ \quad & \frac{-\sum_{k\in U} y_k}{(\sum_{k\in U} 1)^2}\left[\sum_{k\in U} \frac{b_k (I_k^A + (1-I_k^A)I_k^B)}{(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k)} + \sum_{k\in U} (1-b_k)\,\frac{I_k^A}{\pi_k^A} - \sum_{k\in U} 1\right] \\
= \quad & \sum_{k\in U} \frac{(I_k^A + (1-I_k^A)I_k^B)}{\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k} v_k + \sum_{k\in U} \frac{I_k^A}{\pi_k^A} w_k,
\end{aligned}$$

where

$$v_k = \frac{1}{\sum_{k\in U} 1}\left[y_k b_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} 1} b_k\right],$$

and

$$w_k = \frac{1}{\sum_{k\in U} 1}\left[y_k\,(1-b_k) - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} 1}\,(1-b_k)\right].$$

The design variance assuming Poisson sampling for s_B is

$$\begin{aligned}
& \mathrm{Var}\left[\mathrm{E}\left[\sum_{k\in U} \frac{(I_k^A + (1-I_k^A)I_k^B)}{\pi_k^A + (1-\pi_k^A)\rho_k} v_k + \sum_{k\in U} \frac{I_k^A}{\pi_k^A} w_k \,\middle|\, s_A\right]\right] \\
+ \quad & \mathrm{E}\left[\mathrm{Var}\left[\sum_{k\in U} \frac{(I_k^A + (1-I_k^A)I_k^B)}{\pi_k^A + (1-\pi_k^A)\rho_k)} v_k + \sum_{k\in U} \frac{I_k^A}{\pi_k^A} w_k \,\middle|\, s_A\right]\right] \\
= \quad & \mathrm{Var}\left[\sum_{k\in U} \frac{(I_k^A + (1-I_k^A)\rho_k)}{\pi_k^A + (1-\pi_k^A)\rho_k} v_k + \sum_{k\in U} \frac{I_k^A}{\pi_k^A} w_k\right] + \mathrm{E}\left[\sum_{k\in U} \frac{v_k^2 (1-I_k^A)^2 \rho_k (1-\rho_k)}{(\pi_k^A + (1-\pi_k^A)\rho_k)^2}\right] \\
= \quad & \sum_{k,\ell\in U} \left(\pi_{k\ell}^A - \pi_k^A \pi_\ell^A\right) \frac{h_k}{\pi_k^A} \frac{h_\ell}{\pi_\ell^A} + \sum_{k\in U} \frac{v_k^2 (1-\pi_k^A)\rho_k (1-\rho_k)}{(\pi_k^A + (1-\pi_k^A)\rho_k)^2},
\end{aligned} \tag{4.10}$$

where $h_k = \left\{ \left((1-\rho_k) v_k \pi_k^A \right) \left(\pi_k^A + (1-\pi_k^A)\rho_k \right)^{-1} + w_k \right\}$. The variance estimator that estimates the first component of (4.10) from the probability sample using the with-replacement approximation, and estimates the second component of (4.10) from the probability sample s_A assuming Poisson sampling for s_B is:

$$\begin{aligned} & \frac{1}{n_A\left(n_A-1\right)} \sum_{k \in s_A}\left(\frac{\widehat{h}_k}{\pi_k^A / n_A}-\sum_{k \in s_A} \frac{(1-\widehat{\rho}_k) \widehat{v}_k}{\pi_k^A+(1-\pi_k^A) \widehat{\rho}_k}+\frac{\widehat{w}_k}{\pi_k^A}\right)^2 \\ + & \sum_{k \in U} \frac{\widehat{v}_k^2(1-\pi_k^A) \widehat{\rho}_k(1-\widehat{\rho}_k)}{\left(\pi_k^A+(1-\pi_k^A) \widehat{\rho}_k\right)^2} \frac{I_k^A}{\pi_k^A}, \end{aligned} \tag{4.11}$$

where $\widehat{h}_k = \left\{ \left((1-\widehat{\rho}_k) \widehat{v}_k \pi_k^A \right) \left(\pi_k^A + (1-\pi_k^A)\widehat{\rho}_k \right)^{-1} + \widehat{w}_k \right\}$.

$$\widehat{v}_k = \frac{1}{\sum_{k \in s_A} 1/\pi_k^A} \left[y_k b_k - \frac{\sum_{k \in s_A} y_k/\pi_k^A}{\sum_{k \in s_A} 1/\pi_k^A} b_k \right],$$

and

$$\widehat{w}_k = \frac{1}{\sum_{k \in s_A} 1/\pi_k^A} \left[y_k \left(1-b_k\right) - \frac{\sum_{k \in s_A} y_k/\pi_k^A}{\sum_{k \in s_A} 1/\pi_k^A} \left(1-b_k\right) \right].$$

The modified separate estimator for the mean $\widehat{R}_{y,\mathrm{sep,m}}$ is

$$\frac{\psi \sum_{k \in U} y_k b_k I_k^A/\pi_k^A + (1-\psi) \sum_{k \in U} y_k b_k (1-I_k^A) I_k^B/(1-\pi_k^A)\widehat{\rho}_k + \sum_{k \in U} y_k \left(1-b_k\right) I_k^A/\pi_k^A}{\psi \sum_{k \in U} b_k I_k^A/\pi_k^A + (1-\psi) \sum_{k \in U} b_k (1-I_k^A) I_k^B/(1-\pi_k^A)\widehat{\rho}_k + \sum_{k \in U} \left(1-b_k\right) I_k^A/\pi_k^A}, \tag{4.12}$$

where ψ is between 0 and 1. The design variance of the modified separate estimator can also be approximated by Taylor expansion, and derived from the iterated variance by conditioning on the s_A sample first. The Taylor expansion is

$$\sum_{k \in U} \psi \frac{I_k^A}{\pi_k^A} v_k + \sum_{k \in U} (1-\psi) \frac{(1-I_k^A) I_k^B}{(1-\pi_k^A)\widehat{\rho}_k} v_k + \sum_{k \in U} \frac{I_k^A}{\pi_k^A} w_k.$$

The design variance assuming Poisson sampling for s_B is

$$\begin{aligned} & \text{Var}\left[\text{E}\left[\sum_{k\in U}\psi\frac{I_k^A}{\pi_k^A}v_k+\sum_{k\in U}(1-\psi)\frac{(1-I_k^A)I_k^B}{(1-\pi_k^A)\rho_k}v_k+\sum_{k\in U}\frac{I_k^A}{\pi_k^A}w_k\middle| s_A\right]\right] \\ + & \text{E}\left[\text{Var}\left[\sum_{k\in U}\psi\frac{I_k^A}{\pi_k^A}v_k+\sum_{k\in U}(1-\psi)\frac{(1-I_k^A)I_k^B}{(1-\pi_k^A)\rho_k}v_k+\sum_{k\in U}\frac{I_k^A}{\pi_k^A}w_k\middle| s_A\right]\right] \\ = & \sum_{k,\ell\in U}\left(\pi_{k\ell}^A-\pi_k^A\pi_\ell^A\right)\frac{g_k}{\pi_k^A}\frac{g_\ell}{\pi_\ell^A}+\sum_{k\in U}(1-\psi)^2\frac{v_k^2(1-\pi_k^A)\rho_k(1-\rho_k)}{\left((1-\pi_k^A)\rho_k\right)^2}, \end{aligned} \tag{4.13}$$

where $g_k=\left\{\psi v_k+\left(-(1-\psi)v_k\pi_k^A\right)\left((1-\pi_k^A)\right)^{-1}+w_k\right\}$. The variance estimator that estimates the first component of (4.13) from the probability sample using the with-replacement approximation, and estimates the second component of (4.13) from the probability sample s_A assuming Poisson sampling for s_B is:

$$\begin{aligned} & \frac{1}{n_A\left(n_A-1\right)}\sum_{k\in s_A}\left(\frac{\widehat{g}_k}{\pi_k^A/n_A}-\sum_{k\in s_A}\frac{\psi\widehat{v}_k+\left(-(1-\psi)\widehat{v}_k\right)}{(1-\pi_k^A)}+\frac{\widehat{w}_k}{\pi_k^A}\right)^2 \\ + & (1-\psi)^2\sum_{k\in U}\frac{\widehat{v}_k^2(1-\pi_k^A)\widehat{\rho}_k(1-\widehat{\rho}_k)}{\left(\left(1-\pi_k^A\right)\widehat{\rho}_k\right)^2}\frac{I_k^A}{\pi_k^A}, \end{aligned} \tag{4.14}$$

where $\widehat{g}_k=\left\{\psi\widehat{v}_k+\left(-(1-\psi)\widehat{v}_k\pi_k^A\right)\left((1-\pi_k^A)\right)^{-1}+\widehat{w}_k\right\}$.

4.3.3 Simulation experiment

The artificial population for the CCTC simulation consists of 113,549 records. Stratified probability samples of size 1,000 are selected from the whole population. Nonprobability samples of size 4,000 are selected from a sub-population consisting of 74,202 records. The inclusion probability is known for the probability sample and unknown for the nonprobability sample. The estimation methods are evaluated on 22 binary response variables related to a person's behavior or attitude. There are many possible covariates for matching and propensity modeling. We use all covariates in matching the nonprobability sample and probability sample, including education, employment status, income, age, race, region, and metro (binary indicator of metropolitan area). We estimate

the inclusion probability of nonprobability sample ρ_k for all variables using the model

$$\text{logit}(\rho_k) = \beta_0 + \beta_1 \text{age}.$$

We use this very simple model because adding other covariates like race did not improve the prediction performance. From the CCTC simulation experiment, we use 1000 replicates of the probability and nonprobability samples and construct combined and separate estimators of 22 variables, plus a baseline using only probability samples and the original weights. The ψ for separate estimator is fixed at 0.5. Table 4.3 summarizes the percent relative bias for each of the 22 variables. The estimator using only the probability sample is theoretically unbiased. Both the combined and separate estimator incorporate the nonprobability sample while maintaining small relative bias.

Table 4.4 summarizes the effective sample size ratio for each of the 22 variables. In the ratio, mean square error for the baseline is the numerator and mean square error for combined or separate estimator is the denominator. If the ratio is greater than 1, there is a benefit of adding nonprobability sample to the probability sample. For all variables, there is a gain from the addition of the nonprobability sample, but the increase is not a factor of five (from 1000 probability to 1000 probability plus 4000 nonprobability) due to the selection and coverage issues with the nonprobability sample.

We compute the standard error for combined and separate estimator from (4.11) and (4.14). The variance estimator is stable and approximately unbiased. Table 4.5 summarizes the 95% confidence interval coverage for the 22 variables. Almost all confidence intervals using either the separate or combined estimator have close to the 95% nominal coverage. In some cases, like the "Q18-Very Unimportant" variable, the population proportion is very close to zero and confidence intervals for such sparse proportions are known to have poor coverage because the normal approximation is poor.

Table 4.3: Monte Carlo estimates (based on 1000 replicated samples) of percent relative bias of estimators using the probability sample only, combined estimator (4.9), and separate estimator (4.12) for 22 binary variables from the CCTC experiment.

Variable	Probability Sample	Combined Estimator	Separate Estimator
Classical music	0.04	1.91	1.22
In person art experience	0.10	0.37	0.26
Online exhibitions	0.04	1.26	0.68
See play	0.23	0.01	0.11
Live online event	-0.26	1.16	0.54
Want more fun	0.17	-0.91	-0.52
Online kid activities	0.07	2.84	1.52
Celebrate heritage	0.62	1.95	1.49
Community festival	0.05	0.15	0.07
Want more hope	0.04	-0.59	-0.35
Watched TV series or movie	-0.06	-0.27	-0.21
Take art class	-0.47	0.31	-0.03
Q17-Very Unimportant	0.00	-2.26	-1.35
Q17-Unimportant	0.70	-0.81	-0.21
Q17-Neither	0.21	-2.33	-1.23
Q17-Important	0.00	-1.72	-1.00
Q17-Very Important	-0.12	1.74	0.96
Q18-Very Unimportant	0.52	-0.28	-0.29
Q18-Unimportant	-0.59	-4.52	-2.94
Q18-Neither	-0.50	-3.54	-2.24
Q18-Important	0.11	-1.67	-0.91
Q18-Very Important	0.05	1.59	0.94

Table 4.4: Effective sample size ratio (based on 1000 replicated samples) of the combined estimator (4.9), and separate estimator (4.12) for 22 binary variables from the CCTC experiment.

Variable	Combined Estimator	Separate Estimator
Classical music	1.61	1.67
In person art experience	2.83	2.11
Online exhibitions	2.11	2.01
See play	3.26	2.35
Live online event	2.48	2.17
Want more fun	2.23	2.05
Online kid activities	1.11	1.53
Celebrate heritage	2.85	2.07
Community festival	2.91	2.18
Want more hope	2.85	2.21
Watched TV series or movie	2.39	1.97
Take art class	3.23	2.22
Q17-Very Unimportant	2.50	1.90
Q17-Unimportant	2.93	2.09
Q17-Neither	2.31	1.99
Q17-Important	2.32	2.02
Q17-Very Important	1.72	1.88
Q18-Very Unimportant	1.85	1.56
Q18-Unimportant	2.00	1.72
Q18-Neither	1.77	1.69
Q18-Important	2.18	1.99
Q18-Very Important	1.51	1.75

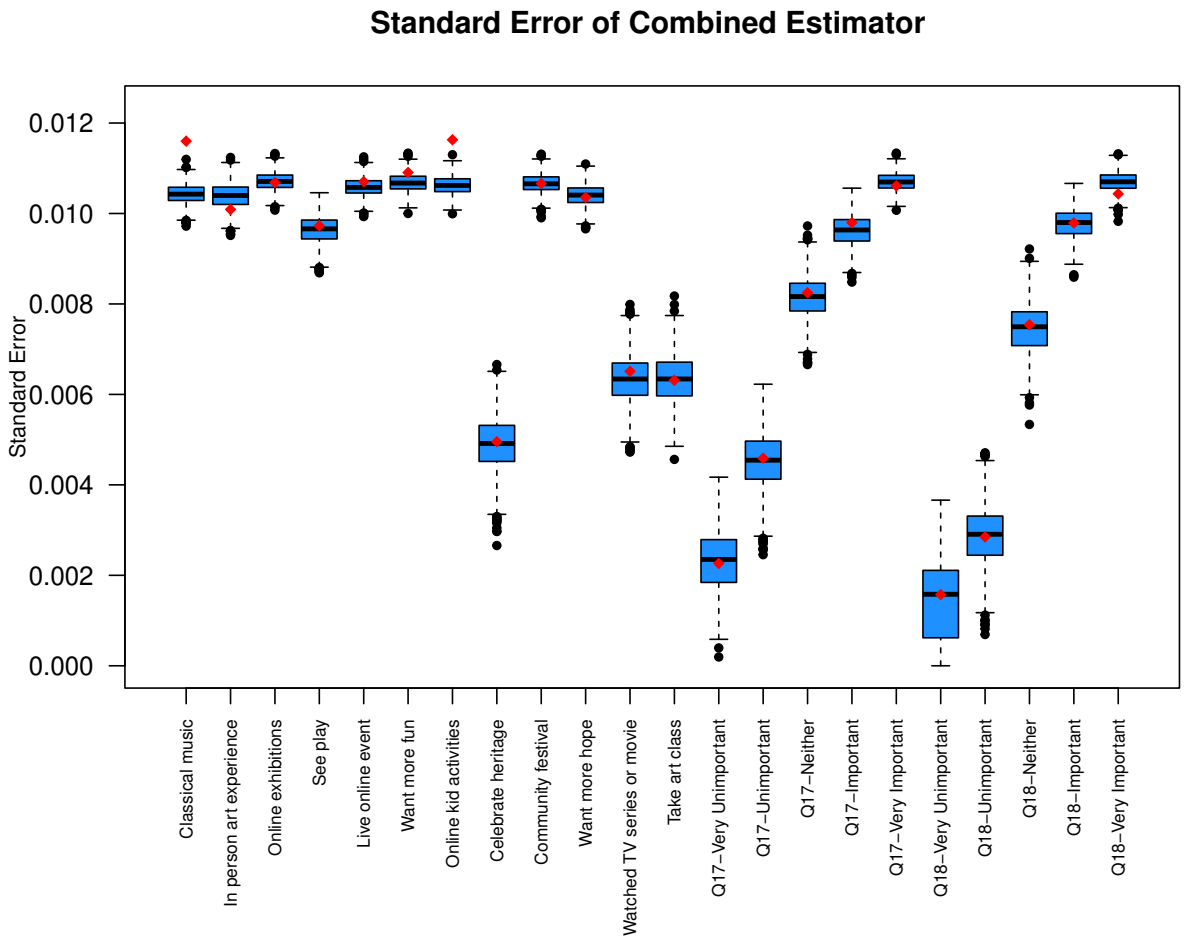

Figure 4.25: Standard error (based on 1000 replicated samples) of the combined estimator for 22 binary variables from the CCTC experiment. The red dots are the true standard error approximated via Monte Carlo.

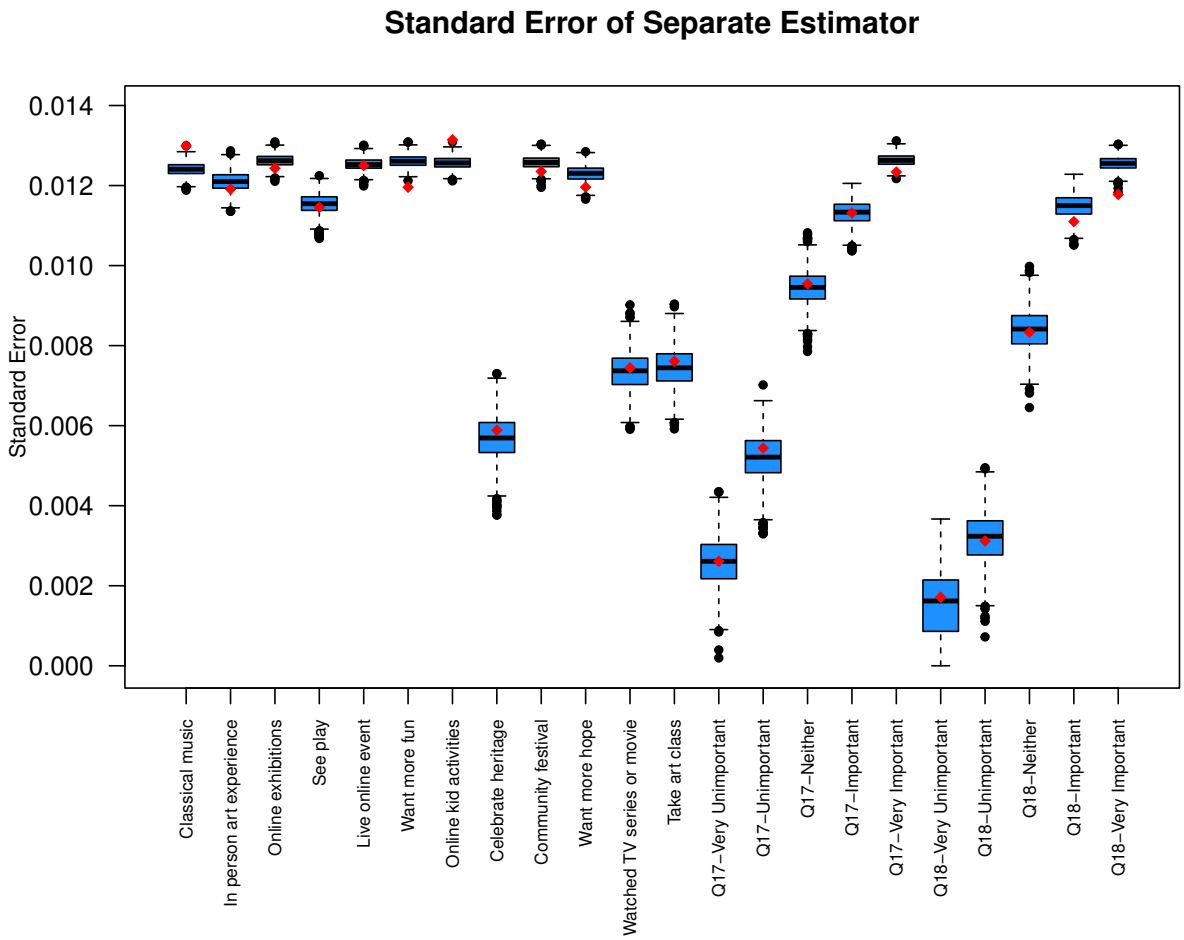

Figure 4.26: Standard error (based on 1000 replicated samples) of the separate estimator for 22 binary variables from the CCTC experiment. The red dots are the true standard error approximated via Monte Carlo.

Table 4.5: 95 percent confidence interval coverage (based on 1000 replicated samples) of the combined estimator (4.9), and separate estimator (4.12) for 22 binary variables from the CCTC experiment.

Variable	Combined Estimator	Separate Estimator
Classical music	85.1	91.7
In person art experience	94.8	94.5
Online exhibitions	89.9	94.0
See play	94.1	94.4
Live online event	92.3	95.5
Want more fun	93.0	96.4
Online kid activities	75.3	90.2
Celebrate heritage	93.5	93.0
Community festival	94.5	95.3
Want more hope	94.5	95.3
Watched TV series or movie	93.8	93.9
Take art class	95.0	94.0
Q17-Very Unimportant	89.5	91.0
Q17-Unimportant	92.7	93.5
Q17-Neither	90.8	94.7
Q17-Important	90.5	93.7
Q17-Very Important	87.1	94.9
Q18-Very Unimportant	73.4	78.9
Q18-Unimportant	89.3	91.5
Q18-Neither	89.8	93.1
Q18-Important	91.7	95.4
Q18-Very Important	85.5	94.4

4.4 Discussion

We give three possible extensions of the dual-frame type estimator combing the probability and nonprobability samples. Doubly-robust estimator fixes the drawback of misspecified probability model by including a prediction model for the response. The dual-frame type estimator applied to the respondent-driven sampling dominates the existing RDS estimator for seeds that are selected with known probability design and a relative small number being recruited. In the setting of incomplete frame of the nonprobability sample, the estimator is also robust by adding the matching step between the probability and nonprobability samples. The dual-frame estimator is robust across different kinds of data set that involves the strict probability sample and the sample with an unknown selection mechanism or inclusion probability.

Chapter 5

Discussion and conclusion

5.1 Summary of contributions

We examined two messy survey problems in this dissertation, imperfect matching in the auxiliary information and inference for surveys incorporating expert judgment. Auxiliary variable in estimation has been widely used in the survey setting, but it assumes the external data source can be matched to one and only one auxiliary record for the kth element in the population. We developed a difference estimator and its variance estimator when the element k is imperfectly matched to the external data source. We also extended the difference estimator to the multiple frame scenario. In the simulation, we demonstrated that the difference estimator under imperfect matching is better than the Horvitz-Thompson type estimator in terms of lower RMSE.

Because of the difficulty and expense of collecting data from a probability sample, there is a growing literature discussing inference for nonprobability samples. The response of interest is available in both probability and nonprobability samples in our scenario of expert judgment sampling. We combine the nonprobability sample with the probability sample to generate the estimates of inclusion probabilities and estimators of population total. The combined estimator dominates the classic probability sample with known weights across a range of characteristics and different judgment behaviors. Theoretical properties including consistency and central limit theorem of the combined estimator are also derived in chapter 3. We extend the estimator in chapter 3 to the doubly-robust combined estimator to avoid the problem of a misspecified probability model. The methodological contributions in chapter 3 are shown to work in a variety of contexts like respondent-driven sampling and incomplete frame of nonprobability samples.

5.2 Directions for further work

In the imperfect matching, three other directions for generalization of the results are (1) allowing the predictor $\mu(\boldsymbol{a}_\ell)$ to be estimated from the sample, (2) allowing the match metric values $m_{k\ell}$ to depend on the sample, and (3) allowing multiple auxiliary databases. The first of these generalizations is standard in the survey literature (see Breidt and Opsomer (2017) for an extensive review), but will be novel in this context due to the uncertain matching. The second generalization is also novel; some of the techniques of Breidt and Opsomer (2008); Dahlke et al. (2013) may be relevant in determining suitable variance estimation strategies.

In the expert judgment inference, future work may involve the generalization of the doubly-robust estimator that allows the different partition of the sample with known or unknown inclusion probability, known or unknown auxiliary information available at the sample level, and known or unknown auxiliary information available at the population level. In the dual-frame approach for combining probability and nonprobability samples, future work may involve the model selection for the propensity model, and the alternate setting of the incomplete frame in section 4.3, in which the nonprobability sample frame gives more complete coverage of the population and the probability sample frame is incomplete.

Bibliography

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–972.

Bankier, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association 81*(396), 1074–1079.

Baraff, A. J., T. H. McCormick, and A. E. Raftery (2016). Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proceedings of the National Academy of Sciences 113*(51), 14668–14673.

Benoit-Bryan, J. and E. Mulrow (2021). Exploring nonprobability methods with simulations from a common data source: culture and community in a time of crisis. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA. American Statistical Association.

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.

Breidt, F. J. and J. D. Opsomer (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics 36*(1), 403–427.

Breidt, F. J. and J. D. Opsomer (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science 32*(2), 190–205.

Breidt, F. J., J. D. Opsomer, and C.-M. Huang (2018). Model-assisted survey estimation with imperfectly matched auxiliary data. In *International Conference of the Thailand Econometrics Society*, pp. 21–35. Springer.

Carpenter, J. R., M. G. Kenward, and S. Vansteelandt (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A (Statistics in Society) 169*(3), 571–584.

Casady, R. J. and M. G. Sirken (1980). A multiplicity estimator for multiple frame sampling. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 601–605. American Statistical Association.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association 115*(532), 2011–2021.

Dahlke, M., F. J. Breidt, J. D. Opsomer, and I. Van Keilegom (2013). Nonparametric endogenous post-stratification estimation. *Statistica Sinica 23*(1), 189–211.

Davidian, M., A. A. Tsiatis, and S. Leon (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science 20*(3), 261–301.

Elliott, M. R. and R. Valliant (2017). Inference for nonprobability samples. *Statistical Science 32*(2), 249–264.

Fellows, I. E. (2019). Respondent-driven sampling and the homophily configuration graph. *Statistics in Medicine 38*(1), 131–150.

Fuller, W. A. (2009). *Sampling Statistics*. New Jersey: Wiley.

Fuller, W. A. and L. F. Burmeister (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 245–249.

Ganesh, N., V. Pineau, A. Chakraborty, and J. M. Dennis (2017). Combining probability and non-probability samples using small area estimation. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 1657–1667. American Statistical Association.

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association 106*(493), 135–146.

Gile, K. J. and M. S. Handcock (2010). Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology 40*(1), 285–327.

Goel, S. and M. J. Salganik (2010). Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences 107*(15), 6743–6747.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*(4), 857–871.

Handcock, M. S., I. E. Fellows, and K. J. Gile (2021). *RDS: Respondent-Driven Sampling*. Los Angeles, CA. R package version 0.9-3.

Hansen, M. H. and W. N. Hurwitz (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics 14*(4), 333–362.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems 44*(2), 174–199.

Heckathorn, D. D. and C. J. Cameron (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology 43*, 101–119.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Isaki, C. and W. A. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association 77*, 89–96.

Kalton, G. and D. W. Anderson (1986). Sampling rare populations. *Journal of the Royal Statistical Society: Series A (General) 149*(1), 65–82.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kim, J. K. and D. Haziza (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica 24*(1), 375–394.

Kim, J. K., S. Park, Y. Chen, and C. Wu (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(3), 941–963.

Kim, J. K. and Z. Wang (2019). Sampling techniques for big data analysis. *International Statistical Review 87*, S177–S191.

Liu, B., L. Stokes, T. Topping, and G. Stunz (2017). Estimation of a total from a population of unknown size and application to estimating recreational red snapper catch in texas. *Journal of Survey Statistics and Methodology 5*(3), 350–371.

Lohr, S. L. (2009). Multiple-frame surveys. In *Handbook of statistics, Sample Surveys: Design, Methods and Applications, Vol. 29A*, pp. 71–88. Amsterdam: North Holland.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology 33*(2), 151–157.

Michaels, S., V. Pineau, B. Reimer, N. Ganesh, and J. M. Dennis (2019). Test of a hybrid method of sampling the lgbt population: Web respondent driven sampling with seeds from a probability sample. *Journal of Official Statistics 35*(4), 731–752.

National Marine Fisheries Service (2015). 2013-2015 Large Pelagics Intercept Survey and Large Pelagics Biological Survey. https://www.st.nmfs.noaa.gov/Assets/recreational/pdf/LPIS\%20Statement\%20of\%20Work\%202013-2015.pdf.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics 33*(2), 659–680.

Rivers, D. (2007). Sampling for web surveys. Paper Prepared for the 2007 Joint Statistical Meetings, Salt Lake City, UT.

Salganik, M. J. and D. D. Heckathorn (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology 34*(1), 193–239.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Singh, A. C. and F. Mecatti (2011). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of Official Statistics 27*(4), 633–650.

Skinner, C. J. and J. N. K. Rao (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association 91*, 349–356.

Thompson, S. K. (2012). *Sampling*. John Wiley & Sons.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology 8*(2), 231–263.

Volz, E. and D. D. Heckathorn (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics 24*(1), 79–97.

Yang, M., N. Ganesh, E. Mulrow, and V. Pineau (2018). Estimation methods for nonprobability samples with a companion probability sample. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 1715–1723. American Statistical Association.

Appendix A

A.1 Likelihood and score function for the with replacement sampling assumption of the nonprobability sample

Suppose the nonprobability sampling design is with replacement sampling, the log-likelihood function for the ρ_k is

$$\ln L(\boldsymbol{\theta}) = \sum_{k \in U \setminus s_A} b_k \ln \rho_k - \sum_{k \in U \setminus s_A} b_k \ln \left(\sum_{k \in U \setminus s_A} \rho_k \right),$$

where b_k is the number of times the kth element in the nonprobability samples. Because the second term of the log-likelihood involves data not in s_A or s_B, it is replaced by the Horvitz-Thompson estimator of its expectation

$$\sum_{k \in U \setminus s_A} b_k \ln \rho_k - \sum_{k \in U \setminus s_A} b_k \ln \left(\sum_{k \in U} \rho_k (1 - \pi_k^A) \frac{I_k^A}{\pi_k^A} \right).$$

Under the logistic regression model of ρ_k, $\text{logit}\left(\rho_k\right) = \boldsymbol{x}_k^\top \boldsymbol{\theta}$, the pseudo log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{k \in U \setminus s_A} b_k \left[\boldsymbol{x}_k^\top \boldsymbol{\theta} - \ln \left\{ 1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right) \right\} \right] - n_B \ln \left\{ \sum_{k \in U} \frac{\exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right)}{1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right)} (1 - \pi_k^A) \frac{I_k^A}{\pi_k^A} \right\},$$

and the score function is

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}) &= \sum_{k \in U \setminus s_A} b_k \left[\boldsymbol{x}_k - \frac{\exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right)}{1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right)} \boldsymbol{x}_k \right] \\ &- n_B \frac{\sum_{k \in U} \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right) \boldsymbol{x}_k \left(1 - \pi_k^A\right) I_k^A / \left\{ 1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right) \right\}^2 \pi_k^A}{\sum_{k \in U} \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right) \left(1 - \pi_k^A\right) I_k^A / \left\{ 1 + \exp\left(\boldsymbol{x}_k^\top \boldsymbol{\theta} \right) \right\} \pi_k^A}. \end{aligned} \tag{A.1}$$

The expected value of (A.1) is

$$\begin{aligned} & \mathrm{E}\left[\mathrm{E}\left[\sum_{k\in U\setminus s_A} b_k\left\{\boldsymbol{x}_k-\frac{\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k}{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)}\right\}\middle| s_A\right]\right] \\ - \quad & n_B\mathrm{E}\left[\frac{\sum_{k\in U}\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k\left(1-\pi_k^A\right)I_k^A/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}^2\pi_k^A}{\sum_{k\in U}\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\left(1-\pi_k^A\right)I_k^A/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}\pi_k^A}\right] = (\mathrm{I})-(\mathrm{II})\,. \end{aligned}$$

For (I),

$$\begin{aligned} & \mathrm{E}\left[\mathrm{E}\left[\sum_{k\in U\setminus s_A} b_k\left\{\boldsymbol{x}_k-\frac{\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k}{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)}\right\}\middle| s_A\right]\right] \\ = \quad & n_B\mathrm{E}\left[\sum_{k\in U\setminus s_A}\frac{\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}}{\sum_{k\in U}\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}}\right] \\ = \quad & n_B\left[\frac{\sum_{k\in U\setminus s_A}\left(1-\pi_k^A\right)\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}^2}{\sum_{k\in U}\left(1-\pi_k^A\right)\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}}\right]+O(n^{-1/2}). \end{aligned}$$

For (II),

$$\begin{aligned} & n_B\mathrm{E}\left[\frac{\sum_{k\in U}\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k\left(1-\pi_k^A\right)I_k^A/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}^2\pi_k^A}{\sum_{k\in U}\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\left(1-\pi_k^A\right)I_k^A/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}\pi_k^A}\right] \\ = \quad & n_B\left[\frac{\sum_{k\in U\setminus s_A}\left(1-\pi_k^A\right)\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\boldsymbol{x}_k/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}^2}{\sum_{k\in U}\left(1-\pi_k^A\right)\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)/\left\{1+\exp\left(\boldsymbol{x}_k^\top\boldsymbol{\theta}\right)\right\}}\right]+O(n^{-1/2}). \end{aligned}$$

The score function is asymptotically unbiased for 0.

A.2 Proof of Lemmas in chapter 3

A.2.1 Lemma 3

Proof. Because $\pi_k = \pi_k^A + (1-\pi_k^A)\rho_k$, by (B1) and $0 \leqslant \rho_k \leqslant 1$, $\min_{k\in U}\pi_k \geqslant \lambda > 0$. By the assumption that s_B follows Poisson sampling, $\pi_{k\ell} = \pi_{k\ell}^A(1-\rho_k)(1-\rho_\ell) + \pi_k^A\rho_\ell(1-\rho_k) + \pi_\ell^A\rho_k(1-\rho_\ell) + \rho_k\rho_\ell$. By (B1) and $0 \leqslant \rho_k \leqslant 1$, $\min_{k,\ell\in U}\pi_{k\ell} \geqslant \lambda^* > 0$. The covariance of the

combined sample for $k \neq \ell$ is

$$\begin{aligned}
\text{Cov}(I_k, I_\ell) &= \text{Cov}\big(I_k^A + (1-I_k^A)I_k^B, I_\ell^A + (1-I_\ell^A)I_\ell^B\big) \\
&= \text{Cov}(I_k^A, I_\ell^A) + \text{Cov}\big(I_k^A, (1-I_\ell^A)I_\ell^B\big) + \text{Cov}\big(I_\ell^A, (1-I_k^A)I_k^B\big) \\
&+ \text{Cov}\big((1-I_k^A)I_k^B, (1-I_\ell^A)I_\ell^B\big) \\
&= \text{Cov}(I_k^A, I_\ell^A) + \text{E}\left[\text{Cov}\left(I_k^A, (1-I_\ell^A)I_\ell^B \middle| s_A\right)\right] \\
&+ \text{Cov}\left(\text{E}\left[I_k^A \middle| s_A\right], \text{E}\left[(1-I_\ell^A)I_\ell^B \middle| s_A\right]\right) + \text{E}\left[\text{Cov}\left(I_\ell^A, (1-I_k^A)I_k^B \middle| s_A\right)\right] \\
&+ \text{Cov}\left(\text{E}\left[I_\ell^A \middle| s_A\right], \text{E}\left[(1-I_k^A)I_k^B \middle| s_A\right]\right) + \text{E}\left[\text{Cov}\left((1-I_k^A)I_k^B, (1-I_\ell^A)I_\ell^B \middle| s_A\right)\right] \\
&+ \text{Cov}\left(\text{E}\left[(1-I_k^A)I_k^B \middle| s_A\right], \text{E}\left[(1-I_\ell^A)I_\ell^B \middle| s_A\right]\right) \\
&= \left(\pi_{k\ell}^A - \pi_k^A\pi_\ell^A\right) - \rho_\ell\left(\pi_{k\ell}^A - \pi_k^A\pi_\ell^A\right) - \rho_k\left(\pi_{k\ell}^A - \pi_k^A\pi_\ell^A\right) + \rho_k\rho_\ell\left(\pi_{k\ell}^A - \pi_k^A\pi_\ell^A\right) \\
&= \left(\pi_{k\ell}^A - \pi_k^A\pi_\ell^A\right)(1-\rho_k)(1-\rho_\ell).
\end{aligned}$$

Then

$$\begin{aligned}
&\limsup_{N\to\infty}(n_A+n_B)\max_{k,\ell\in U:k\neq\ell}|\pi_{k\ell}-\pi_k\pi_\ell| \\
=\ &\limsup_{N\to\infty}(n_A+n_B)\max_{k,\ell\in U:k\neq\ell}\left|\left(\pi_{k\ell}^A-\pi_k^A\pi_\ell^A\right)(1-\rho_k)(1-\rho_\ell)\right| \\
=\ &\limsup_{N\to\infty}\left\{n_A\max_{k,\ell\in U:k\neq\ell}\left|\left(\pi_{k\ell}^A-\pi_k^A\pi_\ell^A\right)\right||(1-\rho_k)(1-\rho_\ell)| \right. \\
+\ &\left. n_B\max_{k,\ell\in U:k\neq\ell}\left|\left(\pi_{k\ell}^A-\pi_k^A\pi_\ell^A\right)\right||(1-\rho_k)(1-\rho_\ell)|\right\},
\end{aligned}$$

which is bounded by (B1) and (B4). □

A.2.2 Lemma 4

Proof. For $k_1, k_2, k_3, k_4 \in D_{4,N}$, the expectation $\mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}\right)\left(I_{k_2} - \pi_{k_2}\right)\left(I_{k_3} - \pi_{k_3}\right)\left(I_{k_4} - \pi_{k_4}\right)\right]$ can be written as

$$
\begin{aligned}
& \mathrm{E}\left[\left\{\left(I_{k_1}^A + (1 - I_{k_1}^A) I_{k_1}^B\right) - \left(\pi_{k_1}^A + (1 - \pi_{k_1}^A)\rho_{k_1}\right)\right\}\right. \\
& \left\{\left(I_{k_2}^A + (1 - I_{k_2}^A) I_{k_2}^B\right) - \left(\pi_{k_2}^A + (1 - \pi_{k_2}^A)\rho_{k_2}\right)\right\} \\
& \left\{\left(I_{k_3}^A + (1 - I_{k_3}^A) I_{k_3}^B\right) - \left(\pi_{k_3}^A + (1 - \pi_{k_3}^A)\rho_{k_3}\right)\right\} \\
& \left.\left\{\left(I_{k_4}^A + (1 - I_{k_4}^A) I_{k_4}^B\right) - \left(\pi_{k_4}^A + (1 - \pi_{k_4}^A)\rho_{k_4}\right)\right\}\right] \\
= \; & \left(1 - \rho_{k_1}\right)\left(1 - \rho_{k_2}\right)\left(1 - \rho_{k_3}\right)\left(1 - \rho_{k_4}\right) \mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}^A\right)\left(I_{k_2} - \pi_{k_2}^A\right)\left(I_{k_3} - \pi_{k_3}^A\right)\left(I_{k_4} - \pi_{k_4}^A\right)\right].
\end{aligned}
$$

By $0 \leqslant \rho_k \leqslant 1$, (B3), (B4), and the Cauchy–Schwarz inequality,

$$
\begin{aligned}
& \lim_{N\to\infty} n^2 \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}\right)\left(I_{k_2} - \pi_{k_2}\right)\left(I_{k_3} - \pi_{k_3}\right)\left(I_{k_4} - \pi_{k_4}\right)\right]\right| \\
= \; & \lim_{N\to\infty} (n_A + n_B)^2 \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\left(1 - \rho_{k_1}\right)\left(1 - \rho_{k_2}\right)\left(1 - \rho_{k_3}\right)\left(1 - \rho_{k_4}\right) \times \right. \\
& \left. \mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}^A\right)\left(I_{k_2} - \pi_{k_2}^A\right)\left(I_{k_3} - \pi_{k_3}^A\right)\left(I_{k_4} - \pi_{k_4}^A\right)\right]\right| \\
= \; & \lim_{N\to\infty} n_A^2 \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\left(1 - \rho_{k_1}\right)\left(1 - \rho_{k_2}\right)\left(1 - \rho_{k_3}\right)\left(1 - \rho_{k_4}\right)\right| \times \\
& \left|\mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}^A\right)\left(I_{k_2} - \pi_{k_2}^A\right)\left(I_{k_3} - \pi_{k_3}^A\right)\left(I_{k_4} - \pi_{k_4}^A\right)\right]\right| \\
+ \; & \lim_{N\to\infty} n_B^2 \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\left(1 - \rho_{k_1}\right)\left(1 - \rho_{k_2}\right)\left(1 - \rho_{k_3}\right)\left(1 - \rho_{k_4}\right)\right| \times \\
& \left|\mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}^A\right)\left(I_{k_2} - \pi_{k_2}^A\right)\left(I_{k_3} - \pi_{k_3}^A\right)\left(I_{k_4} - \pi_{k_4}^A\right)\right]\right| \\
+ \; & \lim_{N\to\infty} 2 n_A n_B \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\left(1 - \rho_{k_1}\right)\left(1 - \rho_{k_2}\right)\left(1 - \rho_{k_3}\right)\left(1 - \rho_{k_4}\right)\right| \times \\
& \left|\mathrm{E}\left[\left(I_{k_1} - \pi_{k_1}^A\right)\left(I_{k_2} - \pi_{k_2}^A\right)\left(I_{k_3} - \pi_{k_3}^A\right)\left(I_{k_4} - \pi_{k_4}^A\right)\right]\right| < \infty.
\end{aligned}
$$

For $k_1, k_2, k_3, k_4 \in D_{4,N}$, the expectation $\mathrm{E}\left[\left(I_{k_1}I_{k_2} - \pi_{k_1k_2}\right)\left(I_{k_3}I_{k_4} - \pi_{k_3k_4}\right)\right]$ can be written as

$$
\begin{aligned}
& \mathrm{E}\left[\left\{\left(I^A_{k_1} + (1-I^A_{k_1})I^B_{k_1}\right)\left(I^A_{k_2} + (1-I^A_{k_2})I^B_{k_2}\right)\right.\right. \\
& \left. - \left(\pi^A_{k_1k_2}\left(1-\rho_{k_1}-\rho_{k_2}+\rho_{k_1}\rho_{k_2}\right) + \pi^A_{k_1}\left(\rho_{k_2}-\rho_{k_1}\rho_{k_2}\right) + \pi^A_{k_2}\left(\rho_{k_1}-\rho_{k_1}\rho_{k_2}\right) + \rho_{k_1}\rho_{k_2}\right)\right\} \\
& \left\{\left(I^A_{k_3} + (1-I^A_{k_3})I^B_{k_3}\right)\left(I^A_{k_4} + (1-I^A_{k_4})I^B_{k_4}\right)\right. \\
& \left.\left. - \left(\pi^A_{k_3k_4}\left(1-\rho_{k_3}-\rho_{k_4}+\rho_{k_3}\rho_{k_4}\right) + \pi^A_{k_3}\left(\rho_{k_4}-\rho_{k_3}\rho_{k_4}\right) + \pi^A_{k_4}\left(\rho_{k_3}-\rho_{k_3}\rho_{k_4}\right) + \rho_{k_3}\rho_{k_4}\right)\right\}\right] \\
= \;& \mathrm{E}\left[\left\{\left(I^A_{k_1}I^A_{k_2} - \pi^A_{k_1k_2}\right)\left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right) + \left(I^A_{k_1} - \pi^A_{k_1}\right)\rho_{k_2}(1-\rho_{k_2})\right.\right. \\
+ \;& \left.\left(I^A_{k_2} - \pi^A_{k_2}\right)\rho_{k_1}(1-\rho_{k_1})\right\} \times \left\{\left(I^A_{k_3}I^A_{k_4} - \pi^A_{k_3k_4}\right)\left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\right. \\
+ \;& \left.\left.\left(I^A_{k_3} - \pi^A_{k_3}\right)\rho_{k_4}(1-\rho_{k_4}) + \left(I^A_{k_4} - \pi^A_{k_4}\right)\rho_{k_3}(1-\rho_{k_3})\right\}\right] \\
= \;& \left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right)\left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\mathrm{E}\left[\left(I^A_{k_1}I^A_{k_2} - \pi^A_{k_1k_2}\right)\left(I^A_{k_3}I^A_{k_4} - \pi^A_{k_3k_4}\right)\right] \\
+ \;& \left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right)\rho_{k_4}(1-\rho_{k_4})\mathrm{E}\left[\left(I^A_{k_1}I^A_{k_2} - \pi^A_{k_1k_2}\right)\left(I^A_{k_3} - \pi^A_{k_3}\right)\right] \\
+ \;& \left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right)\rho_{k_3}(1-\rho_{k_3})\mathrm{E}\left[\left(I^A_{k_1}I^A_{k_2} - \pi^A_{k_1k_2}\right)\left(I^A_{k_4} - \pi^A_{k_4}\right)\right] \\
+ \;& \left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\rho_{k_2}(1-\rho_{k_2})\mathrm{E}\left[\left(I^A_{k_3}I^A_{k_4} - \pi^A_{k_3k_4}\right)\left(I^A_{k_1} - \pi^A_{k_1}\right)\right] \\
+ \;& \left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\rho_{k_1}(1-\rho_{k_1})\mathrm{E}\left[\left(I^A_{k_3}I^A_{k_4} - \pi^A_{k_3k_4}\right)\left(I^A_{k_2} - \pi^A_{k_2}\right)\right] \\
+ \;& \rho_{k_2}(1-\rho_{k_2})\rho_{k_4}(1-\rho_{k_4})\mathrm{E}\left[\left(I^A_{k_1} - \pi^A_{k_1}\right)\left(I^A_{k_3} - \pi^A_{k_3}\right)\right] \\
+ \;& \rho_{k_2}(1-\rho_{k_2})\rho_{k_3}(1-\rho_{k_3})\mathrm{E}\left[\left(I^A_{k_1} - \pi^A_{k_1}\right)\left(I^A_{k_4} - \pi^A_{k_4}\right)\right] \\
+ \;& \rho_{k_1}(1-\rho_{k_1})\rho_{k_4}(1-\rho_{k_4})\mathrm{E}\left[\left(I^A_{k_2} - \pi^A_{k_2}\right)\left(I^A_{k_3} - \pi^A_{k_3}\right)\right] \\
+ \;& \rho_{k_1}(1-\rho_{k_1})\rho_{k_3}(1-\rho_{k_3})\mathrm{E}\left[\left(I^A_{k_2} - \pi^A_{k_2}\right)\left(I^A_{k_4} - \pi^A_{k_4}\right)\right].
\end{aligned}
$$

By (B1),

$$
\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I^A_{k_1} - \pi^A_{k_1}\right)\left(I^A_{k_3} - \pi^A_{k_3}\right)\right]\right| = \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\pi^A_{k_1k_3} - \pi^A_{k_1}\pi^A_{k_3}\right| = 0
$$

Similarly,

$$
\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4 \in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I^A_{k_1} - \pi^A_{k_1}\right)\left(I^A_{k_4} - \pi^A_{k_4}\right)\right]\right| = 0;
$$

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_2}^A - \pi_{k_2}^A\right)\left(I_{k_4}^A - \pi_{k_4}^A\right)\right]\right| = 0;$$

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_2}^A - \pi_{k_2}^A\right)\left(I_{k_3}^A - \pi_{k_3}^A\right)\right]\right| = 0.$$

From the covariance inequality,

$$\left|\mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)\left(I_{k_4}^A - \pi_{k_4}^A\right)\right]\right|^2 \leq \mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)^2\right]\mathrm{E}\left[\left(I_{k_4}^A - \pi_{k_4}^A\right)^2\right].$$

By (B1) and (B3),

$$\begin{aligned}
& \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)\left(I_{k_4}^A - \pi_{k_4}^A\right)\right]\right|^2 \\
\leq & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)^2\right] \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \mathrm{E}\left[\left(I_{k_4}^A - \pi_{k_4}^A\right)^2\right] \\
\leq & 0.
\end{aligned}$$

Hence,

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)\left(I_{k_4}^A - \pi_{k_4}^A\right)\right]\right| = 0.$$

Similarly,

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1}^A I_{k_2}^A - \pi_{k_1k_2}^A\right)\left(I_{k_3}^A - \pi_{k_3}^A\right)\right]\right| = 0;$$

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_3}^A I_{k_4}^A - \pi_{k_3k_4}^A\right)\left(I_{k_1}^A - \pi_{k_1}^A\right)\right]\right| = 0;$$

$$\lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_3}^A I_{k_4}^A - \pi_{k_3k_4}^A\right)\left(I_{k_2}^A - \pi_{k_2}^A\right)\right]\right| = 0.$$

Therefore,

$$
\begin{aligned}
& \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left|\mathrm{E}\left[\left(I_{k_1}I_{k_2}-\pi_{k_1k_2}\right)\left(I_{k_3}I_{k_4}-\pi_{k_3k_4}\right)\right]\right| \\
= & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right)\left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\times \\
& \left|\mathrm{E}\left[\left(I^A_{k_1}I^A_{k_2}-\pi^A_{k_1k_2}\right)\left(I^A_{k_3}I^A_{k_4}-\pi^A_{k_3k_4}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right)\rho_{k_4}(1-\rho_{k_4})\left|\mathrm{E}\left[\left(I^A_{k_1}I^A_{k_2}-\pi^A_{k_1k_2}\right)\left(I^A_{k_3}-\pi^A_{k_3}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left(1-\rho_{k_1}\right)\left(1-\rho_{k_2}\right)\rho_{k_3}(1-\rho_{k_3})\left|\mathrm{E}\left[\left(I^A_{k_1}I^A_{k_2}-\pi^A_{k_1k_2}\right)\left(I^A_{k_4}-\pi^A_{k_4}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\rho_{k_2}(1-\rho_{k_2})\left|\mathrm{E}\left[\left(I^A_{k_3}I^A_{k_4}-\pi^A_{k_3k_4}\right)\left(I^A_{k_1}-\pi^A_{k_1}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \left(1-\rho_{k_3}\right)\left(1-\rho_{k_4}\right)\rho_{k_1}(1-\rho_{k_1})\left|\mathrm{E}\left[\left(I^A_{k_3}I^A_{k_4}-\pi^A_{k_3k_4}\right)\left(I^A_{k_2}-\pi^A_{k_2}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \rho_{k_2}(1-\rho_{k_2})\rho_{k_4}(1-\rho_{k_4})\left|\mathrm{E}\left[\left(I^A_{k_1}-\pi^A_{k_1}\right)\left(I^A_{k_3}-\pi^A_{k_3}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \rho_{k_2}(1-\rho_{k_2})\rho_{k_3}(1-\rho_{k_3})\left|\mathrm{E}\left[\left(I^A_{k_1}-\pi^A_{k_1}\right)\left(I^A_{k_4}-\pi^A_{k_4}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \rho_{k_1}(1-\rho_{k_1})\rho_{k_4}(1-\rho_{k_4})\left|\mathrm{E}\left[\left(I^A_{k_2}-\pi^A_{k_2}\right)\left(I^A_{k_3}-\pi^A_{k_3}\right)\right]\right| \\
+ & \lim_{N\to\infty} \max_{\left(k_1,k_2,k_3,k_4\in D_{4,N}\right)} \rho_{k_1}(1-\rho_{k_1})\rho_{k_3}(1-\rho_{k_3})\left|\mathrm{E}\left[\left(I^A_{k_2}-\pi^A_{k_2}\right)\left(I^A_{k_4}-\pi^A_{k_4}\right)\right]\right| = 0.
\end{aligned}
$$

□

A.3 Assumptions and the proof of the central limit theorem of the combined estimator

(D1) Let $T_{y,N}$ be a sequence of functions of the elements of F_N and let θ_N, $\nu_{0,N}$ be a sequence of constants not depending on F_N. Assume

$$\frac{T_{y,N}-\theta_N}{\sqrt{\nu_{0,N}}} \xrightarrow{\mathcal{L}} N(0,1),$$

where $T_{y,N}=\sum_{k\in U} y_k$.

(D2) Let a sequence of probability samples $s_{A,N}$ selected from F_N such that

$$\left. \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_{y,N}}{\sqrt{\mathrm{Var}\left[\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]\right] \middle| F_N}} \right| F_N \xrightarrow{\mathcal{L}} N(0,1) \text{ a.s.},$$

where

$$\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} | s_A, F_N\right] = \left(\sum_{k \in s_A} \frac{y_k}{\pi_k^A + (1-\pi_k^A)\rho_k} + \sum_{k \in U \setminus s_A} \frac{y_k \rho_k}{\pi_k^A + (1-\pi_k^A)\rho_k} \right),$$

and

$$\mathrm{Var}\left[\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] \middle| F_N\right] = \sum_{k,\ell \in U} \left(\pi_{k\ell}^A - \pi_k^A \pi_\ell^A\right) \frac{y_k^*}{\pi_k^A} \frac{y_\ell^*}{\pi_\ell^A}$$

with $y_k^* = (1-\rho_k)\pi_k^A y_k \left(\pi_k^A + (1-\pi_k^A)\rho_k\right)^{-1}$.

(D3) Assume

$$V_A V_B^{-1} \big| F_N \to \gamma \text{ in probability as } N \to \infty,$$

where γ is a nonzero constant.

Lemma 5. *Assume a sequence of subsamples $s_{B,N}$ selected from the complement of the probability sample and follows Poisson sampling, then*

$$\left. \frac{\widehat{T}_{y,\mathrm{com}} - E\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{Var\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}} \right| s_{A,N}, F_N \xrightarrow{\mathcal{L}} N(0,1) \textit{ a.s.},$$

where

$$Var\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] = \sum_{k \in U} \frac{(1-I_k^A)\rho_k(1-\rho_k)}{(\pi_k^A + (1-\pi_k^A)\rho_k)^2} y_k^2.$$

Proof. Let $r_k = \frac{y_k\left(1-I_k^A\right)\left(I_k^B-\rho_k\right)}{\pi_k^A+(1-\pi_k^A)\rho_k}, k = 1, 2, \cdots, N$. Under Poisson sampling for the $s_{B,N}$ and conditioned on the $s_{A,N}$ and F_N, r_k are independent random variables with mean 0 and variance

$$V_k = \frac{y_k^2\left(1-I_k^A\right)^2}{\left(\pi_k^A+(1-\pi_k^A)\rho_k\right)^2}\rho_k(1-\rho_k).$$

Let

$$B_N^2 = \sum_{k\in U}\frac{y_k^2\left(1-I_k^A\right)^2}{\left(\pi_k^A+(1-\pi_k^A)\rho_k\right)^2}\rho_k(1-\rho_k) = \sum_{k\in U}V_k.$$

From assumptions (B1) and (B2), for some $\delta > 0$,

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{k\in U}\mathrm{E}\left[|r_k|^{2+\delta}\right] \le \limsup_{N\to\infty}\frac{1}{N}\sum_{k\in U}\frac{|y_k|^{2+\delta}}{\lambda^{2+\delta}}\times 4^{2+\delta} < \infty \text{ a.s.},$$

and

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{k\in U}V_k \le \limsup_{N\to\infty}\frac{1}{N}\sum_{k\in U}\frac{y_k^2}{4\lambda^2} < \infty \text{ a.s.}.$$

If y_k does not identically equal to 0, and ρ_k does not identically equal to 0 or 1,

$$\sum_{k\in U}\mathrm{E}\left[|r_k|^{2+\delta}\right] = o\left(B_N^{2+\delta}\right).$$

Then

$$B_N^{-1}\sum_{k\in U}r_k\Bigg| s_{A,N}, F_N \xrightarrow{\mathcal{L}} N(0,1) \text{ a.s.}.$$

by 1.9.2 Corollary in Serfling (1980). □

Proof of Result 3.

Proof. $\mathrm{Var}\left[\widehat{T}_{y,\mathrm{com}}\middle| s_A, F_N\right]$ is design mean square consistent by (B1) and (B2). We have

$$V_B\left(\mathrm{Var}\left[\widehat{T}_{y,\mathrm{com}}\middle| s_A, F_N\right]\right)^{-1}\Bigg| F_N \xrightarrow{P} 1. \tag{A.2}$$

From (A.2), (D2) and Lemma 5,

$$\left.\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}}\right| s_{A,N}, F_N \xrightarrow{\mathcal{L}} N(0,1) \text{ a.s.} \tag{A.3}$$

by Slutsky's theorem.

Let E_p denote the expectation conditioned on F_N. For $d \in \mathbb{R}$,

$$
\begin{aligned}
& P\left(\left(\widehat{T}_{y,\mathrm{com}} - T_y\right)(V_A+V_B)^{-1/2} \leqslant d \middle| F_N\right) \\
= \;& P\left(\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] + \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y \;\leqslant d\sqrt{V_A+V_B} \middle| F_N\right) \\
= \;& P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_B}} \middle| F_N\right) \\
= \;& \mathrm{E}\left[P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}}\right.\right. \\
\leqslant \;& \left.\left. d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_B}} \middle| s_{A,N}, F_N\right)\right] \\
= \;& \mathrm{E}_p\left[\Phi\left(d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} | s_A\right] - T_y}{\sqrt{V_B}}\right)\right] \\
+ \;& \mathrm{E}_p\left[P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_B}} \middle| s_{A,N}\right)\right. \\
& \left. -\Phi\left(d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_B}}\right)\right] \\
\leq \;& \mathrm{E}_p\left[\Phi\left(d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} | s_A\right] - T_y}{\sqrt{V_B}}\right)\right] \\
+ \;& \mathrm{E}_p\left[\left|P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_B}} \middle| s_{A,N}\right)\right.\right. \\
& \left.\left. -\Phi\left(d\sqrt{1+\frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_B}}\right)\right|\right] \\
= \;& (\mathrm{I}) + (\mathrm{II}).
\end{aligned}
$$

Let $w = d\left(1 + V_A V_B^{-1}\right)^{1/2} - \left(\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y\right)(V_B)^{-1/2}$. For (II), because

$$\left| P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant w \middle| s_{A,N} \right) - \Phi(w) \right|$$

is bounded for all w. Hence,

$$\begin{aligned} & \lim_{N\to\infty} \mathrm{E}_p\left[\left| P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A\right]}{\sqrt{V_B}} \leqslant w \middle| s_{A,N} \right) - \Phi(w) \right| \right] \\ = \; & \mathrm{E}_p\left[\lim_{N\to\infty} \left| P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A\right]}{\sqrt{V_B}} \leqslant w \middle| s_{A,N} \right) - \Phi(w) \right| \right] \end{aligned}$$

by the dominated convergence theorem. Because Φ is continuous, by Lemma 3.2 of Rao (1962) and (A.3),

$$\lim_{N\to\infty} \sup_w \left| P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant w \middle| s_{A,N} \right) - \Phi(w) \right| = 0$$

Hence,

$$\begin{aligned} & \mathrm{E}_p\left[\lim_{N\to\infty} \left| P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant w \middle| s_{A,N} \right) - \Phi(w) \right| \right] \\ \leq \; & \mathrm{E}_p\left[\lim_{N\to\infty} \sup_w \left| P\left(\frac{\widehat{T}_{y,\mathrm{com}} - \mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right]}{\sqrt{V_B}} \leqslant w \middle| s_{A,N} \right) - \Phi(w) \right| \right] = 0. \end{aligned}$$

For (I), since Φ is a bounded and continuous function,

$$\mathrm{E}_p\left[\Phi\left(d\sqrt{1 + \frac{V_A}{V_B}} - \frac{\mathrm{E}\left[\widehat{T}_{y,\mathrm{com}} \middle| s_A, F_N\right] - T_y}{\sqrt{V_A}} \frac{\sqrt{V_A}}{\sqrt{V_B}} \right) \right] \to \mathrm{E}\left[\Phi\left(d\sqrt{1+\gamma} - Z\sqrt{\gamma} \right) \right]$$

by assumptions (D2), (D3), and the Portmanteau theorem. Hence,

$$P\left(\frac{\widehat{T}_{y,\text{com}}-T_y}{\sqrt{V_A+V_B}}\leqslant d\middle| F_N\right)\to \mathrm{E}\left[\Phi\left(d\sqrt{1+\gamma}-Z\sqrt{\gamma}\right)\right].$$

Let U be a standard normal random variable independent of Z,

$$\begin{aligned}&\mathrm{E}\left[\Phi\left(d\sqrt{1+\gamma}-Z\sqrt{\gamma}\right)\right]=\mathrm{E}\left[P\left(U\leqslant d\sqrt{1+\gamma}-Z\sqrt{\gamma}\middle| Z\right)\right]\\ =\;& P\left(U+Z\sqrt{\gamma}\leqslant d\sqrt{1+\gamma}\right)=\Phi(d).\end{aligned}$$

Hence,

$$P\left(\frac{\widehat{T}_{y,\text{com}}-T_y}{\sqrt{V_A+V_B}}\leqslant d\middle| F_N\right)\to\Phi(d)\text{ as }N\to\infty.$$

If $\nu_{0,N}\ll V_A,V_B$, $(V_A+V_B)\left(\nu_{0,N}+V_A+V_B\right)^{-1}\xrightarrow{P}1$,

$$\frac{\widehat{T}_{y,\text{com}}-\theta_N}{\sqrt{\nu_{0,N}+V_A+V_B}}\xrightarrow{\mathcal{L}}N(0,1),$$

because

$$\frac{\widehat{T}_{y,\text{com}}-T_y}{\sqrt{V_A+V_B}}\frac{\sqrt{V_A+V_B}}{\sqrt{\nu_{0,N}+V_A+V_B}}+\frac{T_y-\theta_N}{\sqrt{\nu_{0,N}+V_A+V_B}}\frac{\sqrt{\nu_{0,N}}}{\sqrt{\nu_{0,N}}}\xrightarrow{\mathcal{L}}N(0,1).$$

If $\nu_{0,N}$ is not small relative to V_A,V_B, by Theorem 1.3.6 of Fuller (2009), assumption (D1) and the central limit theorem of the combined estimator imply that

$$\frac{\widehat{T}_{y,\text{com}}-\theta_N}{\sqrt{\nu_{0,N}+V_A+V_B}}\xrightarrow{\mathcal{L}}N(0,1).$$

□

A.4 Parametric models for simulated trips and catch

Table A.1: Parameters of the trip model (3.9) within each stratum.

Stratum	$\bar{p}$	$\bar{\lambda}$	η	**Stratum**	$\bar{p}$	$\bar{\lambda}$	η
1	0.61	1.93	4	16	0.79	1.85	1
2	0.33	3.65	5	17	0.38	3.43	2
3	0.49	2.75	5	18	0.61	2.41	1
4	0.37	1.75	4	19	0.87	1.04	1
5	0.83	2.17	4	20	0.94	1.00	1
6	0.54	3.85	7	21	0.26	2.34	4
7	0.24	2.81	9	22	0.08	1.22	4
8	0.47	2.21	8	23	0.22	0.80	7
9	0.70	1.47	2	24	0.44	0.60	4
10	0.86	1.75	4	25	0.76	0.66	3
11	0.36	1.70	2	26	0.80	1.86	2
12	0.49	2.23	3	27	0.42	1.39	2
13	0.43	1.73	4	28	0.75	0.89	3
14	0.71	1.64	4	29	0.68	1.08	4
15	0.80	1.47	4	30	0.72	0.49	4

Table A.2: Parameters of the catch model (3.10) given trips for eleven different catch types.

Catch	θ	a	γ	q
No relation	$\text{Unif}(0, 5)$	0	∞	1
Binary	0.3			
Square root in trips: retention	0.8	0.5	4	1
Linear in trips: retention	0.8	1	4	1
Quadratic in trips: retention	0.8	2	4	1
Square root in trips: moderate	4	0.5	∞	1
Linear in trips: moderate	4	1	∞	1
Quadratic in trips: moderate	4	2	∞	1
Square root in trips: high	8	0.5	∞	0.4
Linear in trips: high	8	1	∞	0.4
Quadratic in trips: high	8	2	∞	0.4