Winter 12-19-2020

Integrating Different Data Sources for Estimation of Total with Unknown Population Size

Zhaoce Liu

Southern Methodist University, zhaocel@smu.edu

INTEGRATING DIFFERENT DATA SOURCES FOR ESTIMATION OF TOTAL WITH UNKNOWN POPULATION SIZE

Approved by:

Dr. S. Lynne Stokes
Professor of Statistical Science, SMU

Dr. Daniel F. Heitjan
Professor in Department of Statistical Science, SMU & Population & Data Sciences, UTSW

Dr. Jing Cao
Associate Professor of Statistical Science, SMU

Dr. Cornelis J. Potgieter
Assistant Professor of Statistical Science, TCU

INTEGRATING DIFFERENT DATA SOURCES FOR ESTIMATION OF TOTAL WITH UNKNOWN POPULATION SIZE

A Dissertation Presented to the Graduate Faculty of the
Dedman College
Southern Methodist University
in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy
with a
Major in Statistics
by
Zhaoce Liu

B.S., Mathematics, Capital Normal University
M.S., Statistics, Sam Houston State University

December 19, 2020

ACKNOWLEDGMENTS

First, I would like to express my my deepest gratitude to my advisor, Dr. Lynne Stokes, for her persistent guidance and confidence in my ability to complete this dissertation. During the journey of pursing my Ph.D. degree, Dr. Stokes is the one who led me into the survey filed, invoked my interest, and shaped my research ability. She provided me with three research topics and encouraged me to attend many conferences, which enriched my research experience and built my confidence. I also appreciate her patience on my writing. Without her careful editing, I would not have completed this work.

Second, I would like to thank my other committee members: Dr. Heitjan, Dr. Cao and Dr. Potgieter for their valuable comments and advice on this dissertation. It was also my honor to be enrolled in their classes, where I learned so much from their expertise. I also want to express my thanks to our department and all other faculty members, who made me feel as warm as home.

Finally, I dedicate this dissertation to my parents, for their endless love and support. It was their constant encouragement that gave me the strength to get through this journey.

Liu, Zhaoce

B.S., Mathematics, Capital Normal University
M.S., Statistics, Sam Houston State University

Integrating Different Data Sources for Estimation of Total with Unknown Population Size

Advisor: Dr. S. Lynne Stokes

Doctor of Philosophy degree conferred December 19, 2020

Dissertation completed August 21, 2020

Probability sampling has served as the gold-standard in survey practice for many decades. However, as many new data collection methods become available, it is possible to improve the quality and efficiency of traditional survey practices by integrating different sample sources. Web-based surveys from the so-called opt-in panels are one type of non-probability sample that becoming popular these years. They often come with large sample sizes to yield efficient estimates, but selection bias may compromise the generalizability of results to the broader population.

Our motivating example is a survey conducted by National Marine Fisheries Service (NMFS), which collects data to estimate catch of recreational anglers. Currently, the samples are from two surveys, a mail survey measuring effort (# of trips made in a given area) and an intercept survey measuring catch per unit effort (# of fish per trip by species). The samples are combined to provide an estimate of total catch. However, NMFS is experimenting with alternative data collection procedures that use self-reports submitted by anglers via electronic devices, such as cell phones. The self-reports are from a non-probability sample of anglers and may not be accurate. The objective is to improve the quality and speed of estimation, and/or to reduce cost.

This dissertation consists two pieces of research that are both related to this problem. The first part of this dissertation is about finding the sampling design for the current estimators to meet the desired precision. Currently, the estimators proposed by Liu et al.

(2017) treat the self reports as auxiliary data to the sample of intercepts, so they are not used directly in estimation. The estimators' precision depend on several factors, including reporting rate, the accuracy and representativeness of reported counts, and the size of the dockside sample. We develop the R package *OptimalFisheryDesign* to compare the estimation precision of the new estimators, investigate the effects of different factors, and find the corresponding optimal designs for various implementations of the pilot survey.

The second part of this dissertation is to investigate whether or not better estimators of catch can be developed by treating the large sample of voluntary reports as actual data, rather than simply as auxiliary information to improve estimates from the dockside sample. To integrate the non-probability sample and the probability samples, we modify and evaluate two different weighting approaches proposed by Robbins et al. (2015): joint weighting and disjoint weighting. In the joint weighting approach, the samples are only representative when combined as one sample, while in disjoint weighting each sample is weighted to be individually representative of the population, and then averaged.

In addition to PSA, we propose a new method called Adaptive Propensity Score Adjustment (APSA). The method serves as an indicator of whether the propensity score model correctly predicts the selection probability. It can also reduce the selection bias by detecting and dropping part of the non-probability sample whose selection mechanism can not be explained by the model. Both the jackknife and bootstrap methods are proposed and examined for variance estimation.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure Page

LIST OF TABLES

To my parents.

CHAPTER 1

Introduction

1.1. Overview of Non-probability Sampling

Probability sampling has served as the gold-standard in survey practice for many decades (Stephan, 1948; Frankel and Frankel, 1987). The essential component of probability sampling is its random selection mechanism with known probability of selection for each sample unit. This allows valid inferences about parameters from the target population. A probability sample, when analyzed properly, accurately represents the population and thus avoids selection bias. Classical approaches in survey sampling are well discussed in Fuller (2011) and Lohr (2019). However, traditional survey practice has recently been supplemented with new survey data collection methods (Couper, 2005). Many of these have an unknown data generating process, and must be analyzed differently than probability samples. These samples are called non-probability sample.

Compared to traditional survey practice, non-probability sampling has the advantage that its recruitment process is more efficient and less costly, so that larger sample sizes are feasible. It may also require less time to deploy or obtain responses. However, the disadvantage of non-probability sampling is that the data generating process is unknown and thus may not produce samples that can be made to represent the target population well. This causes great challenges for making defensible inferences about the population. The "Summary Report of the AAPOR Task Force on Non-probability Sampling" (Baker et al., 2013), which was commissioned by the American Association of Public Opinion

Research (AAPOR) Executive Council, summarizes three major issues that arise when analyzing non-probability samples: (i) Large exclusion bias: The population accessible to recruitment is likely to be a small and unrepresentative portion of the target population of interest; (ii) Selection bias: participants from the non-probability sample may not be representative, even for the population that was exposed to recruitment. (iii) Non-participation bias: even though a non-probability sample may have a large sample size compared to a probability sample, the participation rates (conditional on being recruited for the study) are often low. Thus, one should be cautious when analyzing non-probability samples.

The remarkable usage of non-probability-sample-based surveys can be tracked back to the 1936 election polls. In pre-election polls, the Literary Digest magazine distributed 10 million straw poll ballots, among which 2.3 million were collected. They revealed that Alf Landon would win by a landslide over Franklin Roosevelt. However, the magazine ignored the fact that the respondents consisted mostly of telephone owners and the magazine's readers, which only represented the middle-to-upper income class of the society at that time and thus introduced severe selection bias into the sample. As a result, the Literary Digest made an erroneous prediction. However, it is notable that Wang et al. (2015) correctly predicted the 2012 presidential election result based on XBOX gaming players, which demonstrates the potential value of non-probability samples if correctly used.

In addition to making inference solely based on the non-probability samples, there are many opportunities as well as challenges of developing methods and frameworks to combine different data sources to assist in estimation. This is known as data integration. This area is facing challenges and opportunities in developing methods and frameworks, as the data sources differ in their quality and suitability for answering research questions, and many of the inexpensive data sources provide non-probability samples (Lohr et al., 2017).

1.2. Different Types of Non-probability Samples

Non-probability samples have long been acceptable in various research fields. For example, in medical research, the recruitment of patients due to their accessibility and availability results in non-probability samples that provide useful clinical findings from experiments. For estimation of population characteristics, however, adoption of non-probability samples for scientific inquiry has been viewed with skepticism. Lately, however, researchers have begun to investigate whether such samples might provide useful information. Due to different selection mechanisms, there are many types of non-probability samples. We briefly summarize some of them here.

1. Mall intercepts: As the name suggests, the potential sample units are intercepted in shopping malls or other public spaces. It is an efficient way of collecting samples that is widely used in marketing research. The mall intercept process involves stopping shoppers, screening them for qualifying characteristics, and then either conducting the interview or inviting the sample units to the appointed research facility for a complete interview. The recruitment process could be random or by some systematic selection mechanism like stopping every tenth shopper for the interviewer encounter.

2. River sampling: In river sampling, the responses are collected from website visitors via online banners, ads, promotions, offers and invitations. The website visitor who clicks on the survey link will be asked several screening questions and finally routed to a survey based on their answers. Once the participant completes the survey, or has been screened out of the survey, they may never be tracked again.

3. Network sampling: Also known as snowball sampling, whose future subjects are collected by referral from the existing sample units. The initial sample units can also be collected by convenience. As illustrated by the name, the sample group grows like a rolling snowball. This sampling method is often used for studying hidden

populations, such as drug users or sex workers, which are hard to reach. Estimation from such samples can be affected by severe bias and are hard to generalize to a broader population.

4. Quota sampling: In quota sampling, the sample units are selected to match the population, in terms of proportions of certain characteristics, such as age and gender. The purpose of quota sampling is to make the sample mimic the population in terms of the given characteristics, with the goal of reducing the bias.

5. Volunteer Panel: The volunteer panel is common in areas such as psychology, social and medical research. Participants who are willing to take part in the study during a certain time period voluntarily join the panel. When a panel is recruited online, its size is often large, with thousands or even millions of members. However, the number of active panel members in any certain time period is limited due to low response rate.

Among the various types of non-probability samples, web-based samples from volunteer panels have become most popular in recent years (Grana et al., 2014; Schonlau et al., 2017). These panels consist of volunteers who are willing to participate and are enlisted through various convenient methods. A web survey is a simple way of getting access to many respondents from the target population. When collecting data through the internet, interviewers are no longer needed. Questionnaires can be distributed at very low cost and thus the survey can be launched easily. In our application, the volunteer panel collects fishing trip information from fishing boat captains via cell phone or satellite devices, which forms a non-probability sample.

1.3. Research Objectives

This research focuses on integrating a non-probability sample with a probability sample for estimation of population total. It was motivated by a data collection method used to collect information from the population of recreational anglers by the National Oceanic and Atmospheric Administration (NOAA). NOAA has been pushed by the state fish and game agencies, who carry out data collection from anglers in their marine waters, to allow new sources of data be used as part of the fish catch estimation process called MRIP (Marine Recreational Information Program). In particular, the anglers themselves are interested in providing data on their catch via technologies such as cell phones or satellite devices on their vessels. Such voluntarily submitted data can be regarded as a non-probability sample. One approach to use these data in estimation is to think of it as part of a capture recapture system, where the self-reported data are the capture phase and a probability sample collected from the dockside in-person survey (MRIP) is the recapture phase. With this use, the self reports serve as the auxiliary information to the recapture sample and thus are not used directly in estimation.

There are two research objectives in this dissertation that are both related to this problem. The first objective is about the sample design for the capture-recapture view of the aggregated data set. Several state fish and game agencies are interested in how to design such a data collection method, using a cell phone app as the self-report mechanism. A frequently encountered question by these agencies is to determine how large the dockside in-person survey sample is required for adequate precision given specific self-reporting rates they hope to achieve. To address this question, we developed the R package *OptimalFisheryDesign* to investigate the effects of various factors on the estimation precision and to find the most cost effective designs for implementation.

The second research objective is to determine whether or not better estimators of catch can be developed by treating the large sample of voluntary reports as actual data, rather than simply as auxiliary information to improve estimates from the dockside sample. To incorporate the self-reported sample directly into estimation of total, one way is to estimate the inclusion probability for the self-reported sample by propensity score adjustment (PSA) (Valliant, 2019). If this step produces accurate estimation of the selection probability, the self-reported sample will have properties of a probability sample and thus can be weighted to make inference on the target population.

The rest of this dissertation is organized as follows. In Chapter 2, we review the research background and estimators of total proposed by Liu et al. (2017). Chapter 3 investigates the first research objective, which is to find the optimal design for the currently used capture-recapture approach that uses the volunteer data only as auxiliary information. Chapter 4 investigates the second research objective, which is to develop inference methods for a sample composed of both the probability and non-probability samples.

CHAPTER 2

Background

2.1. Motivating Example

The National Marine Fisheries Service (NMFS) is responsible for collecting data on catch by the recreational fishing sector. Their data collection operations are known as the Marine Recreational Information Program (MRIP). Data from this program are an input to models that monitor the health of many of the nation's fisheries. For the last 30 years, these data have come from a pair of surveys, one to measure effort (# of trips made) and one to measure catch per unit effort (CPUE=# of fish caught per trip), denoted as $\bar{y}$. Effort is estimated from a retrospective household survey that collects data directly from anglers on their trips in the previous two months, called a wave. CPUE ($\bar{y}$) is estimated from a dockside in-person survey where sampling units are defined by time units and geography. The interviewers must have the expertise to identify all species encountered, and so are technicians/biologists supplied by state fish and game agencies for the coastal states. Final estimates of catch are obtained by multiplying estimated effort and CPUE ($\bar{y}$), which is repeated for each species, geography, and wave.

This system has recently had increasing demands. Some states are interested in in-season management of species to prevent overfishing, requiring quicker processing of data. Scientists need finer geographic resolution to monitor the impact on fishing stock from events, such as the 2010 Deep-water Horizon oil spill. Current sample sizes cannot produce sufficient estimation precision for small geographies and short time intervals.

Increasing the sample sizes sufficiently using the current two-probability-sample methodology is prohibitively expensive in many cases.

Because managers and scientists want greater precision, various states are experimenting with alternative data collection systems. One is the electronic logbook (ELB); this approach allows anglers to self-report their own effort and catch in (near) real time, usually by cell phone or other communication devices. The dockside intercept survey is still used, but its data are combined with the volunteered ELB reports. If these catch reports could capture effort, then estimates would be available sooner since the retrospective household survey might be eliminated. If the catch was accurately reported at a high rate, then estimation of CPUE ($\bar{y}$) might be improved.

Since the self-reported ELB data is not a probability sample, analysts cannot simply substitute them for the household survey in the current estimation system. Instead, they need new estimation methods to ensure that the catch estimates are scientifically defensible. The most common approach has been to use the ELB data as auxiliary information only. This leads to the approach mentioned in Chapter 1, where the reported and dockside intercept samples are viewed as a capture and recapture. Then an estimator similar to the Lincoln-Peterson index (Le Cren, 1965) is used to estimate not the number of trips (N), but the total catch (t_y) from the trips. Several variations of the estimator have been proposed (Liu et al., 2017, Breidt et al. 2018). Though these estimators are consistent for total catch, their precision depends on several factors, including reporting rate, accuracy of the reported count, representativeness of the reporting sample and the size of the dockside intercept sample. The sample size of the dockside intercept sample is under the control of the samplers. However, the features of the reporting sample are less controllable, since they rely on the voluntary participation of anglers in the fishery.

The agencies that consider changing to such an electronic reporting system usually have two questions. The first one is whether the dockside intercept sample size they already have is adequate for their precision needs if used alongside an ELB system. If not, they could increase it or try to influence the number or quality of reports made by anglers to improve precision. For example, they can use outreach programs to educate anglers about the purpose and importance of the ELB program to try to increase reporting rate or quality.

The second question is whether a more efficient estimator can be constructed by using the reported catch directly as data, rather than simply as auxiliary information. Currently, the estimators of total catch (Liu et al., 2017; Breidt et al., 2018) using the volunteer reports treat the reported catch and trips as auxiliary variables only. That is, their content is not taken as data, but rather simply as an auxiliary variable that can be used as a ratio estimator to help improve the estimator made from the dockside survey. However, the reports actually include the variable whose total is being estimated from the dockside survey (catch) and the number of catch reports is much larger among the reporting sample than in the dockside sample. If the self reports can be correctly weighted to estimate the total, the new estimator will take the advantage of a larger sample size and thus has better precision.

2.2. Review of Current Estimators of Total

2.2.1. Capture-Recapture Model

Capture-recapture methods are powerful for population size estimation. Suppose we are interested in estimating the total number of fishes, say N, in a lake. Two catch attempts are made to estimate this quantity. The first attempt selects a sample of n_1 fish, which are marked and released. The second sample of n_2 fish is selected randomly from the same population, and it is found that m of them were previously caught and marked in the first catch. Under the assumption that proportions of marked fish are the same on average in the second sample and the population, we can equate the two proportions:

$$\frac{n_1}{N} = \frac{m}{n_2}. \tag{2.1}$$

This gives the classical estimator, referred to as the Lincoln-Petersen index due to the pioneering work of two ecologists (Le Cren 1965):

$$\hat{N} = \frac{n_1 n_2}{m}. \tag{2.2}$$

This estimator is also the maximum likelihood estimator (MLE) under a hyper-geometric model.

2.2.2. Estimators of Total Catch from Electronic Reports

In our application, the self-reported sample and the dockside intercept sample can be viewed as coming from a capture recapture study. Like the example introduced in Subsection 1.4.1, the self-reported trips can be treated as the capture sample and the dockside intercept sample as the recapture sample. Following previous notation, let N denote the unknown population size, which is the number of recreational fishing trips. Let y denote the fish catch from each trip and X the associated auxiliary information, which is a vector including covariates like the sailing duration, date, and the number of passengers on the boat. Our goal is to estimate the total fish catch $t_y = \sum_{i=1}^{N} y_i$ over a population of unknown size, rather than to estimate the population size itself.

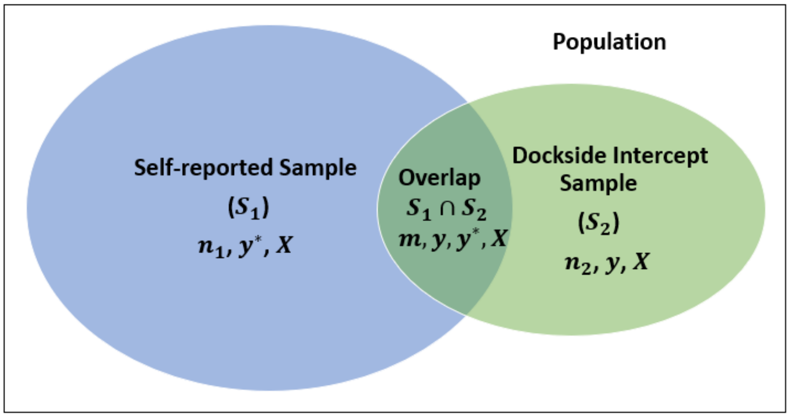


Figure 2.1: Visualization of the two sample sources: the self-reported sample (ELB) and the dockside intercept sample (MRIP).

Figure 2.1 shows our target population and the two sample sources. The self-reported sample (S_1) consists of n_1 trips, each containing the self-reported catch y^* and auxiliary information X. Due to measurement error, y^* may be different from y, and the actual catch y is unobservable for the self-reported trips unless it was also selected into the dockside intercept sample (S_2). S_2 consists of n_2 trips, each containing the actual catch y and auxiliary information X. The overlap between the two samples contains m matched trips with observable y, y^* and auxiliary information X.

Pollock et al. (1994) previously considered this estimation problem, but in his application y^* was not available for S_1. He proposed estimating t_y by

$$\hat{t}_{yp\_SRS} = \hat{N}\bar{y}_{S_2}, \tag{2.3}$$

where $\hat{N} = \frac{n_1 n_2}{m}$ and $\bar{y}_{S_2}$ is the sample average of the $y's$ from the recaptured units. This estimator is appropriate when S_2 is a simple random sample. It can be adapted when S_2 has a complex design, as we will discuss subsequently.

We now review three estimators of total catch proposed by Liu et al. (2017). One is a generalization of $\hat{t}_{yp\_SRS}$ which can be used when the intercept sample has a complex design. The other two are also based on the capture-recapture idea, but make use of self-reported catch available from the probability samples. Following their notation, let r_i be the reporting indicator ($r_i = 1$ if the i^{th} trip is included in S_1, 0 otherwise) and w_i be the inverse of the selection probability, or sampling weight, for the i^{th} trip in S_2. Then $\hat{n}_1 = \sum_{i \in s_2} w_i r_i$, $\hat{p_1} = \sum_{i \in s_2} w_i r_i / \sum_{i \in s_2} w_i$ and $\widehat{\bar{y}} = \sum_{i \in s_2} w_i y_i / \sum_{i \in s_2} w_i$ are estimators of number of reports (n_1), reporting rate (p_1) and CPUE ($\bar{y}$), all made from S_2.

The generalization of Pollock's estimator for a complex design has the form of a ratio estimator with auxiliary variable r_i and ratio $B_p = \frac{t_y}{n_1}$:

$$\hat{t}_{yp} = \frac{n_1}{\hat{p}_1} \widehat{\bar{y}} = n_1 \frac{\hat{t}_y}{\hat{n}_1}. \tag{2.4}$$

The second estimator is also a ratio estimator with auxiliary variable $r_i y_i^*$ and ratio $B_c = \frac{t_y}{t_{y^*}}$:

$$\hat{t}_{yc} = t_{y^*} \frac{\sum_{i \in s_2} w_i y_i}{\sum_{i \in s_2} w_i r_i y_i^*} = t_{y^*} \frac{\hat{t}_y}{\hat{t}_{y^*}}, \tag{2.5}$$

where $t_{y^*} = \sum_{i \in S_1} y^*$ is the total catch from self-reported trips. This estimator can be thought of as making a multiplication adjustment to account for unreported catch.

The third estimator is adapted from an optimal linear combination of the previous two, $\hat{t}_{yp}$ and $\hat{t}_{yc}$:

$$\hat{t}_{MR} = (1 - w)\hat{t}_{yp} + w\hat{t}_{yc}, \tag{2.6}$$

where $0 \leq w \leq$1. Ideally, w would be chosen so that it minimizes the variance. This is a special case of a multivariate ratio estimator proposed by Olkin (1958). In practice, w needs to be estimated from the sample. Under the simplified situation where the recapture sample (intercept dockside sample S_2) is a simple random sample and the self reports are accurate ($y = y^*$), the variance minimizing value of w is $w_{SRS} = t_{y^*}/t_y$. This can be estimated by $\hat{w}_{SRS} = \frac{t_{y^*}}{\hat{t}_{yc}}$. By substituting this value for w in equation 2.6, even when S_2 is not a simple random sample, we obtain:

$$\hat{t}_{y2} = t_{y^*} + \frac{n_1}{\hat{n}_1}(\hat{t}_y - \hat{t}_{y^*}). \tag{2.7}$$

$\hat{t}_{y2}$ is similar to a regression estimator (Section 4, Lohr, 2019), and can be thought of as making an additive adjustment to account for unreported catch.

All the proposed estimators use the self-reported catch as auxiliary data only. Liu et al. (2017) showed that these estimators will improve the estimation precision when the auxiliary variables are highly correlated with the variable of interest. This corresponds to the situation when the reporting rate is high and the reported catch is accurate. Besides possibly improved precision, all three estimators have another advantage over estimators of total made from the intercept sample alone. As noted earlier, the intercept sample can have substantial under-coverage in areas where a substantial fraction of angling sites are inaccessible private sites. Thus the Horvitz-Thompson estimator of total made from the probability sample alone is biased downward. However, this is not necessarily true for the proposed estimators. Specifically, if the average catch, average reported catch, and reporting rate are the same for trips ending in public and private sites, then all three estimators are approximately unbiased (Stokes et al., 2019a). This is the reason these estimators are preferred for all areas where there is a non-negligible fraction of trips ending in private (inaccessible) fishing sites.

CHAPTER 3

OptimalFisheryDesign: An R package for Fishery Sampling Designs

In this Chapter, we introduce the R package *OptimalFisheryDesign* to aid in the sample design as we described in Chapter 2. It designs samples for obtaining data to use for the total catch estimation using estimators proposed by Liu et al. (2017). The package helps the analysts investigate:

1. the estimation precision of the three estimators under different combinations of the dockside intercept sample size and reporting rate,
2. the trade-offs between the dockside intercept sample size and the reporting rate of the three estimators with respect to the estimation precision,
3. the optimal sampling designs of the three estimators under budget constraints.

3.1. The General Approach to Investigate the Fishery Sampling Design

Our goal is to provide a tool for the analyst to help determine the required sample size or reporting rate to obtain a specified precision requirement. More specifically, we want to find the minimum dockside intercept sample size to achieve a specified precision given reporting rate or vice versa, for each of the three estimators $\hat{t}_{yp}$, $\hat{t}_{yc}$ and $\hat{t}_{y2}$. Since there are different combinations of the dockside intercept sample size and reporting rate for the same precision, we are also interested in finding the most cost-effective designs under budget constraints.

To find such designs, we rely on the percent standard error (PSE) expressions for the three estimators. The PSE is a measure of the estimator's precision that is used by NMFS for sample design. According to NMFS, a PSE value greater than 50% indicates a very imprecise estimate.

The PSEs of the three estimators are defined by the ratio of their corresponding standard deviations over the total catch (t_y). To approximate their standard deviations when the intercept sample is from a complex design, we adjust their standard deviations when the dockside intercept sample is a simple random sample by the design effect ($deff$), which is an input made by the user. The design effect is a factor that summaries the effects of various complexities in the sample design, especially those of clustering and stratification (Kish, 1995). Plausible ranges of the design effect are likely to be known by organizations who regularly use complex designs for their dockside intercept sample. Liu et al. (2017) provided expressions (A.9) - (A.12) in Appendix A to approximate the three estimators' standard deviations when the dockside intercept sample is a simple random sample (SRS). Hence, we have:

$$\begin{aligned} PSE(\hat{t}_{yp}) &\approx \frac{\sqrt{Var(\hat{t}_{yp})/deff}}{t_y} \\ &\approx \sqrt{\frac{(1-\frac{n_2}{N})}{n_2/deff}\left\{CV_y^2 + (1+\frac{1}{p_1}) - 2\frac{\bar{y}_1}{\bar{y}}\right\}}, \end{aligned} \tag{3.1}$$

$$\begin{aligned} PSE(\hat{t}_{yc}) &\approx \frac{\sqrt{Var(\hat{t}_{yc})/deff}}{t_y} \\ &\approx \sqrt{\frac{(1-\frac{n_2}{N})}{n_2/deff}\left\{CV_y^2 + (1+\frac{1}{p_1}) - 2\frac{\bar{y}_1}{\bar{y}} + \frac{CV_{1y^*}^2}{p_1} - 2\frac{\bar{y}_1}{\bar{y}}R_{1,yy^*}CV_{1y}CV_{1y^*}\right\}}, \end{aligned} \tag{3.2}$$

$$
\begin{aligned}
PSE(\hat{t}_{y2}) &\approx \frac{\sqrt{Var(\hat{t}_{y2})/deff}}{t_y} \qquad (3.3)\\
&\approx \sqrt{\frac{(1-\frac{n_2}{N})}{n_2/deff}\left\{CV_y^2 + (1+\frac{1}{p_1}) - 2\frac{\bar{y}_1}{\bar{y}} + p_1\frac{\bar{y}_1^*}{\bar{y}}CV_{1y^*}(\frac{\bar{y}_1^*}{\bar{y}}CV_{1y^*} - 2\frac{\bar{y}_1}{\bar{y}}R_{1,yy^*}CV_{1y})\right\}},
\end{aligned}
$$

where $\bar{y}$ is the overall mean catch, $\bar{y}_1$ is the mean catch among the self-reported trips, $\bar{y}_1^*$ is the mean of y^* in the reporting sample, CV_y is the coefficient of variance of y, CV_{1y} and CV_{1y^*} are the coefficients of variation of y and y^* in the reporting sample, and R_{1,yy^*} is the correlation coefficient of y and y^* in the reporting sample.

It is clear from Equations (3.1), (3.2) and (3.3) that the precision of all three estimators can be improved by either increasing the dockside intercept sample size (n_2) or the reporting rate (p_1). If one or the other of the two factors is fixed, the three expressions can help the analyst determine whether the other can be increased sufficiently to meet precision requirements, given the other parameters affecting the PSE.

The other parameters that affect the PSE may not always be available to the sample designer. Hence, we clarify what parameters would be required for the designer to provide and what defaults can be used when some parameters are difficult to predict. These parameters can be classified into two groups.

The first group includes parameters describing the catch distribution ($\bar{y}$, CV_y) and the design effect ($deff$) of the dockside intercept sample. These are parameters that the analyst who has data from past intercept sampling is likely to be able to approximate or predict, so they are required to be provided.

The second group includes parameters related to reporting characteristics, which are $\bar{y}_1$, CV_{1y}, $\bar{y}_1^*$, CV_{1y^*} and R_{1,yy^*}. These parameters may be more difficult to predict for a sample designer that has no experience with an ELB system. For example, If a user has not implemented an electronic reporting system before, he or she will not be able to provide estimates about how complete and representative is the self-reported sample

$(p_1,\bar{y}_1, CV_{1y}, \bar{y}_1^*, CV_{1y^*})$ and how accurate the reports are (R_{1,yy^*}). In that case, the user may provide estimates from a ELB implementation in another state, or use default values. Our approach does not require that these parameters be specified, but rather will provide default settings when they are not. The default settings are: 1) the self reports are accurate ($R_{1,yy^*} = 1$), 2) when none of the $\bar{y}_1$, CV_{1y}, $\bar{y}_1^*$ and CV_{1y^*} can be specified, the self reports are assumed representative of the population $(\bar{y} = \bar{y}_1 = \bar{y}_1^*$, $CV_y = CV_{1y} = CV_{1y^*})$, 3) when either one pair of ($\bar{y}_1$, CV_{1y}) or ($\bar{y}_1^*$, CV_{1y^*}) cannot be specified, the user can choose one of the two following measurement error models to calculate one missing pair from the other:

3a) The classical measurement error model (CME) (Carroll et al., 2006):

$$y^* = y + e,$$

where $e \sim (0, \alpha S_y^2)$, with y and e independent. Under CME, $R_{1,yy^*} = \frac{1}{\sqrt{1+\alpha}}$, $CV_{1y} = CV_{1y^*}/\sqrt{1+\alpha}$.

3b) The Berkson model (Berkson, 1950):

$$y = y^* + e,$$

where $e \sim (0, \beta S_{y^*}^2)$, with y and e independent. Under the Berkson model, $R_{1,yy^*} = \frac{1}{\sqrt{1+\beta}}$, $CV_{1y} = CV_{1y^*}\sqrt{1+\beta}$.

So far, we have discussed all parameters in Equations (3.1), (3.2) and (3.3) based on the given dockside intercept sample size (n_2) and the reporting rate (p_1). When the dockside intercept sample size (n_2) changes, the second set of parameters except R_{1,yy^*} can be affected and we assume the first set of parameters and R_{1,yy^*} do not change. However, when the reporting rate (p_1) changes, the first set of parameters can't be affected but the second set of parameters can change dynamically. This is because the change of the reporting rate (p_1) will affect the representativeness and accuracy of the self-reported sample (Clearly, as p_1 approaches 1, $CV_{1y} = CV_y$, so the parameters are related.). To model this dynamic relationship, we make the following assumptions so $\bar{y}_1$, CV_{1y}, $\bar{y}_1^*$, CV_{1y^*} can be specified as a function of p_1 (Appendix A):

1. when the reporting rate (p_1) increases, the new self-reports are representative of the anglers who did not report before,
2. when the reporting rate (p_1) decreases, the losing self reports are representative of the current self-reported sample,
3. the accuracy of the self-reported sample (R_{1,yy^*}) does not change.

3.2. The R package *OptimalFisheryDesign*

3.2.1. Overview

The package *OptimalFisheryDesign* (available on *Github: Charlieliu004/Fishery-OptimalDesign*) aims to investigate and compare the effect of the dockside intercept sample size (n_2) and the reporting rate (p_1) on the precision of the three estimators: $\hat{t}_{yp}$, $\hat{t}_{yc}$ and $\hat{t}_{y2}$. The package can help the user understand what would be required to design an efficient ELB system for a particular species. For example, it can provide the PSE for a catch estimate for a specified intercept sample size over a range of reporting rates. This can help the user determine whether or not obtaining a desired PSE is feasible, given the known characteristics of the intercept sample. It can also be used to select a cost-effective designs for a given precision and cost.

The package can be used to inform the user either for planning a pilot study (pre-ELB implementation stage) or to adjust an ongoing ELB operation. The user can input what is known about the required parameters of catch and reported catch, and the package will use modeling and default settings described in Subsection 3.1 to supply the rest. Clearly, the more accurate the information the user can provide, the more realistic the outcomes provided will be.

3.2.2. Description of the Functions and Their Inputs

The package has 5 functions, with the key inputs summarized in Table 3.1. The first function *CV_population* is used to calculate the mean and CV of the catch for the dockside intercept sample and the self-reported sample. The function *CV_population* requires three inputs: the percentage of fishing trips with non-zero landings for the whole sample, and the mean and variance of catch from such trips.

The remaining four functions are *InterceptSampleSize*, *ReportingRate*, *Tradeoff* and *OptimalDesign*. All require at a minimum that the user supply the parameters of the catch distribution ($\bar{y}$ and CV_y) for the species that can be calculated by *CV_population* and the design effect of the dockside intercept sample. The characteristics of the self-reported sample are also needed by the four functions as optional inputs, or default values will be assigned. To describe the self-reported sample, the current reporting rate (p_1) and at least one pair of parameters: ($\bar{y}_1,\ CV_{1y}$) or ($\bar{y}_1^*,\ CV_{1y^*}$) are required at minimum, with the remaining parameters supplied by default if not specified. Next, we describe the four functions and their specific inputs.

The function *InterceptSampleSize* investigates the effect of the dockside intercept sample size on the PSE of the three estimators. It requires the additional inputs of the target reporting rate ($target\_p1$) and target precision ($target\_PSE$). For a given reporting rate, the function displays PSE ($\hat{t}_{yp}$), PSE ($\hat{t}_{yc}$) and PSE ($\hat{t}_{y2}$) as functions of the dockside intercept sample size. The function also provides the dockside intercept sample sizes required for the three estimators to achieve their target precision. If the current dockside intercept sample size (n_obs) is specified as the optional input, the three estimators' PSEs for the current intercept sample size and reporting rate will also be specified.

The function *ReportingRate* investigates the effect of the reporting rate on the PSE for the three estimators. It requires the additional inputs of the the target dockside intercept sample size ($target\_n2$) and target precision ($target\_PSE$). Under the target dockside intercept sample size, the function displays PSE ($\hat{t}_{yp}$), PSE ($\hat{t}_{yc}$) and PSE ($\hat{t}_{y2}$) as functions of the reporting rate. The function also specifies the reporting rates required for the three estimators to achieve their target precisions.

The function *Tradeoff* investigates the trade-offs between the dockside intercept sample size and reporting rate for each estimator to achieve the target PSE. It requires the additional input of the target precision ($target\_PSE$). Under the target PSE, the function displays the required dockside intercept sample sizes for the three estimators as functions of the reporting rate. If the target dockside intercept sample size ($target\_n2$) is specified as the optional input, the function will specify the required reporting rates for the three estimators to achieve their target precisions.

The function *OptimalDesign* provides the optimal sampling designs that achieve the smallest PSE for each of the three estimators. It requires the additional input of the cost ratio ($cost\_ratio$) and budget ($RelBudget$). The cost ratio is defined as the cost of increasing one percent reporting rate over the cost of recruiting one dockside intercept sample unit, and the budget is defined in terms of the largest possible dockside sample size that could be collected. Given the cost ratio and budget, the function displays PSE ($\hat{t}_{yp}$), PSE ($\hat{t}_{yc}$) and PSE ($\hat{t}_{y2}$) as functions of the reporting rate. If the current dockside intercept sample size is specified as the optional input, the optimal designs of the three estimators will be determined based on the current design.

Table 3.1: Outline of the key arguments for the package *OptimalFisheryDesign*.

Argument	Description	Corresponding Statistics	Function
Landings_pct	Percentage of fishing trips with non-zero landings		*CV_population*
Landings_mean	Mean of catch among fishing trips with non-zero landings		
Landings_var	Variance of catch among fishing trips with non-zero landings		
Mean_dockside	Mean fish catch of the dockside intercept sample	$\bar{y}$	*InterceptSampleSize* *ReportingRate* *Tradeoff* *OptimalDesign*
CVy	CV of the dockside intercept sample	CVy	
p1_obs	Current reporting rate	p_1	
Mean_report	Mean fish catch of the self-reported sample	$\bar{y}_1$	
CVy_report	CV of the self-reported sample	CV_{1y}	
Mean_report_s	Mean fish catch of the self-reported sample containing measurement error	$\bar{y}_1^*$	
CVy_report_s	CV of the self-reported sample containing measurement error	CV_{1y^*}	
deff	Design effect of the dockside intercept sample	$deff$	
R	Correlation coefficient of the actual fish catch and its self-reported value among the self-reported sample	R_{1,yy^*}	
type	Types of measurement error model. “CME” refers to the classical measurement model, “Berkson” refers to the Berkson model		
n_obs	Current dockside intercept sample size	n_2	*InterceptSampleSize* *OptimalDesign*
target_p1	Desired reporting rate	p_1	*InterceptSampleSize*
target_n2	Desired dockside intercept sample size	n_2	*ReportingRate* *Tradeoff*
target_PSE	Desired estimation precision		*InterceptSampleSize* *ReportingRate* *Tradeoff*
cost_ratio	Relative cost for increasing one percent reporting rate in terms of increasing the number of dockside intercept trips		*OptimalDesign*
RelBudget	Budget in terms of the most possible dockside intercept sample size that could be collected under the current budget		*OptimalDesign*

3.3. Case Study

In this section, we illustrate the usage of the *FisheryOptimalDesign* package by examining a case study on the estimation of the total Red Snapper catch from charter boats in Alabama. In this example, the input parameters are estimated from an ELB experiment conducted in Alabama, which includes both the dockside intercept sample and the self-reported sample. We first describe that data, and then explain how the package could be used for planning the most cost effective designs.

In our example, the dockside intercept sample was collected by NOAA's Marine Recreational Information Program (MRIP), which interviewed anglers during dockside creel surveys, selected according to a complex sample design. The frame of the design consisted of locations crossed with time blocks. The time blocks were stratified by weekday and weekend, while the locations were selected based on a probability proportional to size (PPS) design. The size was measured by "pressure", which was meant to capture the average number of anglers using a particular site in past years. Data about catch/discard counts of different fish species and number of anglers were collected from every vessel intercepted during sampled shifts and locations. Vessel registration numbers were also recorded and used, along with day and time, to identify matches to the self-reported trips. In this example, the dockside intercept sample contained 211 charter trips. We assume the design effect was 1.4, based on the estimate from a similar survey.

The self-reported sample was collected by the Gulf of Mexico Charter Boat E-logbook Project (ELB), which allowed captains to self report their fishing trip information. In our example, the ELB sample contained 1628 self-reported trips. We used a reporting rate of $p_1 = 11\%$ since that was the estimate from the sample. Among the 1628 self-reported trips, 24 were matched to trips from the MRIP sample. Following previous notation, let y and y^* be the total red snapper catch (harvest) of a trip from the MRIP and ELB sample, respectively. Based on the 24 matched trips, the correlation coefficient of y and y^* was estimated to be $R_{1,yy^*} = 0.85$.

Table 3.2 lists the summary statistics for the trips with non-zero fish catch from the MRIP and ELB sample. The statistics listed under “Overlap” were calculated from the matched trips. Table 3.3 lists the total recreational red snapper catch estimated by the three estimators and their corresponding PSEs.

Table 3.2: Descriptive statistics of the recreational Red Snapper catch by Charter boat from the MRIP Sample and ELB Sample.

	MRIP	Overlap	ELB
Percentage of the trips with landings	0.46	0.7	0.94
Mean of catch among the trips with landings	14.62	15.36	16.5
Variance of catch among the trips with landings	113.71	73.5	101.29

Table 3.3: Three estimates of the total recreational Red Snapper catch and their corresponding PSEs.

	Estimate of Total Catch	PSE $(\%)$
$\hat{t}_{yp}$	$69,191$	22.8
$\hat{t}_{yc}$	$42,112$	29.4
$\hat{t}_{y2}$	$57,600$	22.5

The package *FisheryOptimalDesign* can be used to understand how the precision of the three estimators would be improved by either increasing the dockside intercept sample size or the reporting rate, and what are the most cost-effective designs for the three estimators given a specified budget.

3.3.1. Install the *FisheryOptimalDesign* Package

To install the package *FisheryOptimalDesign* from Github into R, we can use the "*install_github*" function in the "*devtools*" package:

```
R > install.packages("devtools")
```

```
R > library(devtools)
R > install_github("Charlieliu004/FisheryOptimalDesign")
R > library("FisheryOptimalDesign")
```

3.3.2. Usage of the Function “*CV_population*”

To estimate $\bar{y}, CVy$ from the MRIP sample, $\bar{y}_1,\ CV_{1y}$ from the matched sample and $\bar{y}_1^*, CV_{1y^*}$ from the ELB sample, we apply the “*CV_population*” function by inputting the statistics listed in Table 3.2:

```
R > #MRIP
R > CV_population(0.46,14.62,113.71)
R > #ELB_est
R > CV_population(0.73,15.15,80.56)
R > #ELB
R > CV_population(0.94,16.51,101.29)
```

The outputs from the function “*CV_population*” are listed in Table 3.4, where the mean catch under “MRIP”, “Overlap” and “ELB” are estimates of $\bar{y}$, $\bar{y}_1$ and $\bar{y}_1^*$, respectively. The CVs under “MRIP”, “Overlap” and “ELB” are estimates of CV_y, CV_{1y} and CV_{1y^*}, respectively.

Table 3.4: Summary statistics of the recreational Red Snapper catch by Charter boat from the MRIP sample and ELB sample.

	MRIP	Overlap	ELB
Mean catch	6.55	15.52	11.06
CV	1.57	0.68	0.92

From Table 3.4, we see that the mean catch from the self reported sample is higher than that of the intercept sample. This is typical for the self-reported catch, as anglers may feel less motivated to report low or no catch.

3.3.3. Usage of the Function "*InterceptSampleSize*"

Suppose we want to reduce the PSE to 16% for all three estimators by increasing the intercept sample size (n_2) with the current reporting rate. To find the required sample sizes for the three estimators, we use the “*InterceptSampleSize*” function with the following inputs:

```
R > InterceptSampleSize(CVy = 1.57,Mean_dockside = 6.55,
+                        target_p1 = 0.11,target_PSE = 16,
+                        Mean_report = 15.52,CVy_report = 0.68,
+                        Mean_report_s = 11.06,CVy_report_s = 0.92,
+                        p1_obs = 0.11, R = 0.85,deff  =1.4)
```

The “*InterceptSampleSize*” function produces Figure 3.1, which displays PSE ($\hat{t}_{yp}$), PSE ($\hat{t}_{yc}$) and PSE ($\hat{t}_{y2}$) as functions of the dockside intercept sample size (n_2) for the reporting rate $p_1 = 0.11$. From Figure 3.1, the PSEs of all three estimators decrease as the sample size increases. All PSE curves in Figure 3.1 are steep when the sample size is small, which indicates the efficiency of the dockside intercept sample in reducing the PSE in this situation. In addition, PSE ($\hat{t}_{yp}$) and PSE ($\hat{t}_{y2}$) are very close over the range of the sample size, while PSE ($\hat{t}_{yc}$) is always higher than the other two. This shows the inefficiency of $\hat{t}_{yc}$ compared to $\hat{t}_{yp}$ and $\hat{t}_{y2}$: to achieve the target PSE of 16%, the dockside intercept sample size should be increased from 211 to 427, 710 and 416 for $\hat{t}_{yp}$, $\hat{t}_{yc}$ and $\hat{t}_{y2}$, respectively.

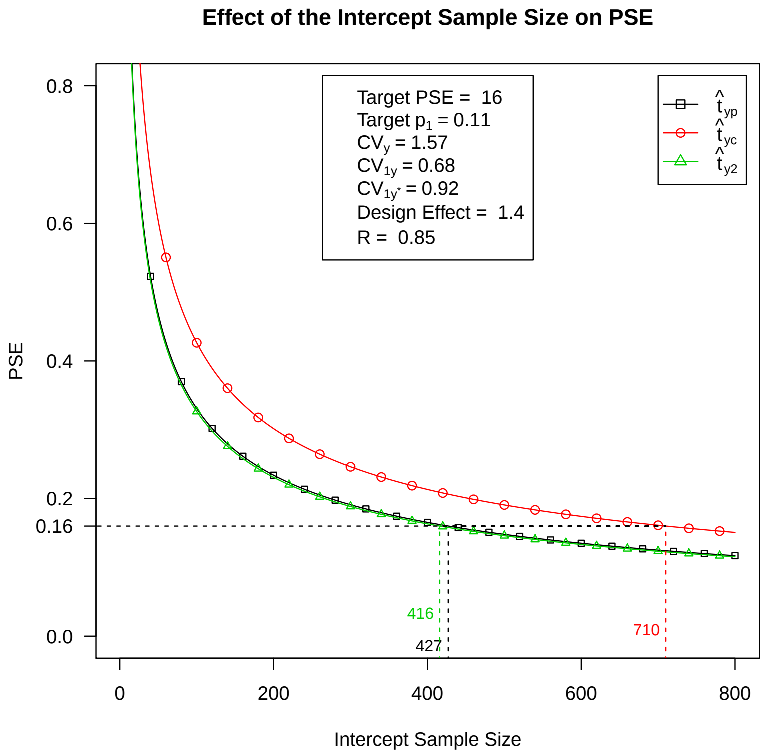


Figure 3.1: Effect of the dockside intercept sample size (n_2) on PSE for total recreational Red Snapper catch by Charter boat.

3.3.4. Usage of the Function "*ReportingRate*"

Suppose we want to reduce the PSE to 16% for all three estimators by increasing the reporting rate (p_1) of the ELB sample. To find the required reporting rates for the three estimators, we use the "*ReportingRate*" function with the following inputs:

```
R > ReportingRate(CVy = 1.57,Mean_dockside = 6.55,
+                 target_n2 = 211,target_PSE = 16,
+                 Mean_report = 15.52,CVy_report = 0.68,
+                 Mean_report_s = 11.06,CVy_report_s = 0.92,
+                 p1_obs = 0.11, R = 0.85,deff  =1.4)
```

The "*ReportingRate*" function produces Figure 3.2, which displays PSE ($\hat{t}_{yp}$), PSE ($\hat{t}_{yc}$) and PSE ($\hat{t}_{y2}$) as functions of the reporting rate (p_1) for the current dockside intercept sample size $n_2 = 211$. Figure 3.2 shows the PSEs of all the three estimators decrease as the reporting rate (p_1) increases, as we would expect. When the reporting rate is low ($< 25\%$), $PSE(\hat{t}_{yp})$ and $PSE(\hat{t}_{y2})$ are very close , while $PSE(\hat{t}_{yc})$ is higher. This indicates $\hat{t}_{yc}$ is less efficient than $\hat{t}_{yp}$ and $\hat{t}_{y2}$ under this scenario: to achieve the target PSE of 16% with the same sample size, the reporting rate must be increased from 11% to 32%, 47% and 27% for $\hat{t}_{yp}$, $\hat{t}_{yc}$ and $\hat{t}_{y2}$, respectively.

Figure 3.2 also expresses similar information as Figure 3.1: increasing the reporting rate when it is low is useful for reducing the PSE. However, it becomes less useful when the reporting rate is high.

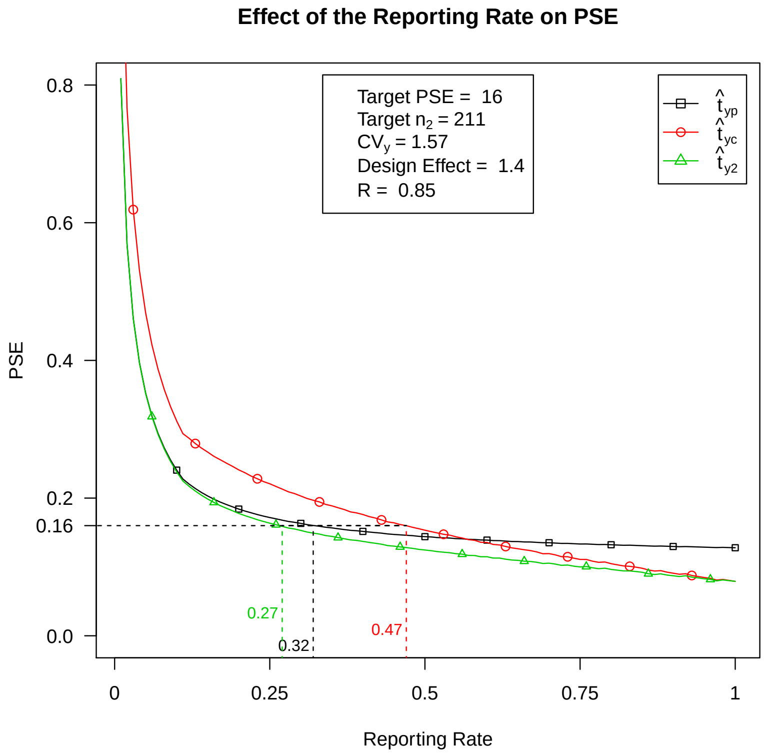


Figure 3.2: Effect of the ELB sample's reporting rate (p_1) on PSE for total recreational Red Snapper catch by Charter boat.

3.3.5. Usage of the Function "*Tradeoff*"

As we showed previously, the PSE can be reduced by either increasing the dockside intercept sample size (n_2) or increasing the reporting rate (p_1) of the ELB sample. To compare the efficiency of the two strategies, we use the "*Tradeoff*" function with the following inputs:

```
R > Tradeoff(CVy = 1.57,Mean_dockside = 6.55,
+           target_n2 = 211, target_PSE = 16,
+           Mean_report = 15.52,CVy_report = 0.68,
+           Mean_report_s = 11.06,CVy_report_s = 0.92,
+           p1_obs = 0.11, R = 0.85,deff  =1.4)
```

The "*Tradeoff*" function produces Figure 3.3, which displays the dockside intercept sample size (n_2) as functions of the ELB sample's reporting rate (p_1) for the three estimators to achieve the target PSE. From Figure 3.3, all three curves are steep when the reporting rate is low. This indicates that by increasing the ELB sample's reporting rate can greatly reduce the required dockside intercept sample sizes for the three estimators. We also notice that $\hat{t}_{yp}$, unlike the other two estimators, does not gain much additional value in precision when reporting rate increases beyond a certain point. This is likely because the estimator does not use the reported catch, but only the number of reported trips in estimation. However, it is still unclear which point on the curve costs least and thus be the most cost-effective designs. This is because the cost ratio as we defined in Subsection 3.2.2 has not been considered.

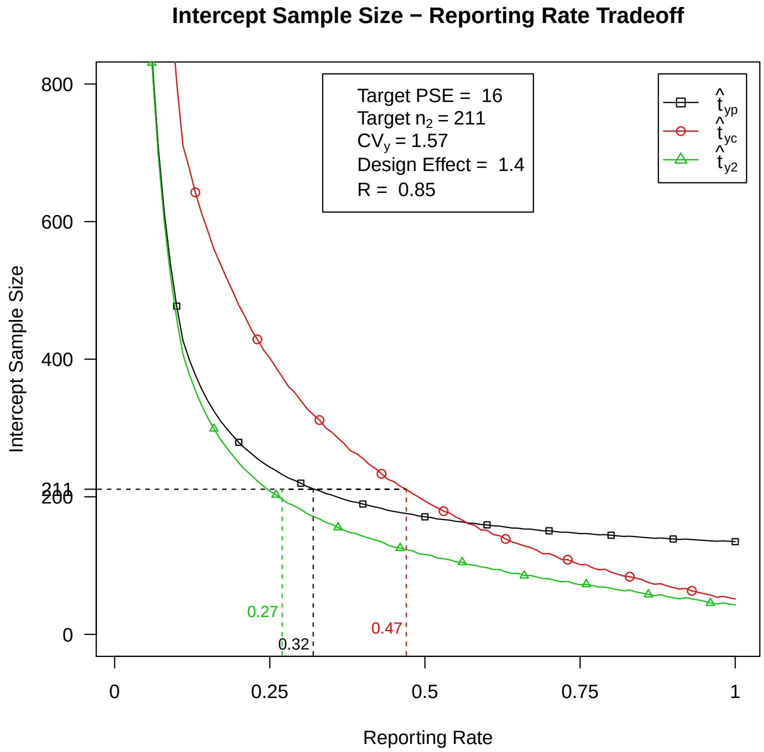


Figure 3.3: Relationship between the dockside intercept sample size (n_2) and the ELB sample's reporting rate (p_1) for total recreational Red Snapper catch by Charter boat.

3.3.6. Usage of the Function "*OptimalDesign*"

Suppose we want to find the most cost-effective designs for all three estimators to achieve the target PSE of 16%, and the cost ratio as we defined in Subsection 3.2.2 is 15. Suppose the budget is either sufficient for increasing the dockside intercept sample units from 211 to 391 or increasing the reporting rate from 11% to 23%. Under this cost ratio and budget, the agency cannot afford to reach the target PSE by purely increasing the dockside intercept sample size (n_2) or the ELB sample's reporting rate (p_1) for all three estimators, as demonstrated by Figure 3.1 and Figure 3.2. In this case, the "*OptimalDesign*" function could provide the most cost-effective sampling design with the following inputs:

```
R > OptimalDesign(CVy = 1.57,Mean_dockside = 6.55,
+                   cost_ratio = 15, RelBudget = 180,
+                   Mean_report = 15.52,CVy_report = 0.68,
+                   Mean_report_s = 11.06,CVy_report_s = 0.92,
+                   n_obs = 211, p1_obs = 0.11, R = 0.85,deff = 1.4)
```

The “*OptimalDesign*” function produces Figure 3.4, which displays PSE ($\hat{t}_{yp}$), PSE ($\hat{t}_{yc}$) and PSE ($\hat{t}_{y2}$) for all possible combinations of the dockside intercept sample size (n_2) and the reporting rate (p_1) of the ELB sample given the budget and cost ratio. Figure 3.4 also provides the optimal sampling designs for the three estimators in the top legend.

The optimal design for $\hat{t}_{yp}$ to reach the target PSE of 16% is by increasing the dockside intercept sample size from 211 to 316, and increasing the reporting rate from 11% to 16%. Similar to $\hat{t}_{yp}$, the optimal design for $\hat{t}_{y2}$ to reach the target PSE is by increasing the dockside intercept sample size from 211 to 301, and increasing the reporting rate from 11% to 17%. However, the optimal design for $\hat{t}_{yc}$ can only reach a PSE of 21%, which is by increasing the dockside intercept sample size from 211 to 331, and increasing the reporting rate from 11% to 15%.

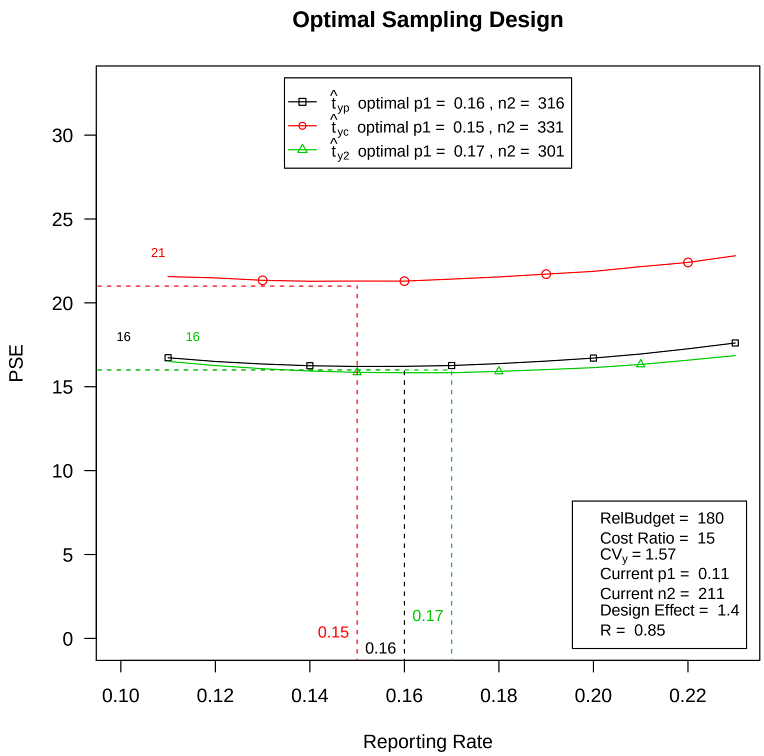


Figure 3.4: Optimal sampling design for total recreational Red Snapper catch by Charter boat.

So far, the total cost for this new design is: $211 + 11 \times 15 + 180 = 556$. The "*OptimalDesign*" function also provides the optimal sampling designs that ignore the current design under this total cost for the three estimators with the following inputs:

```
R > OptimalDesign(CVy = 1.57,Mean_dockside = 6.55,
+                  cost_ratio = 20, RelBudget = 556,
+                  Mean_report = 15.52,CVy_report = 0.68,
+                  Mean_report_s = 11.06,CVy_report_s = 0.92,
+                  n_obs = NULL, p1_obs = 0.11,
+                  R = 0.85,deff = 1.4)
```

The "*OptimalDesign*" function produces Figure 3.5. The optimal designs for the three estimators are provided in the top legend, which are similar to the designs provided in Figure 3.4. Under the given budget and cost ratio, the optimal design for $\hat{t}_{yp}$ can reach the PSE of 16%, with the dockside intercept sample size of 331 and the reporting rate of 15%. The optimal design for $\hat{t}_{y2}$ can reach the PSE of 16%, with the dockside intercept sample size of 316 and the reporting rate of 16%. However, the optimal design for $\hat{t}_{yc}$ can only reach the PSE of 21%, with the dockside intercept sample size of 346 and the reporting rate of 14%.

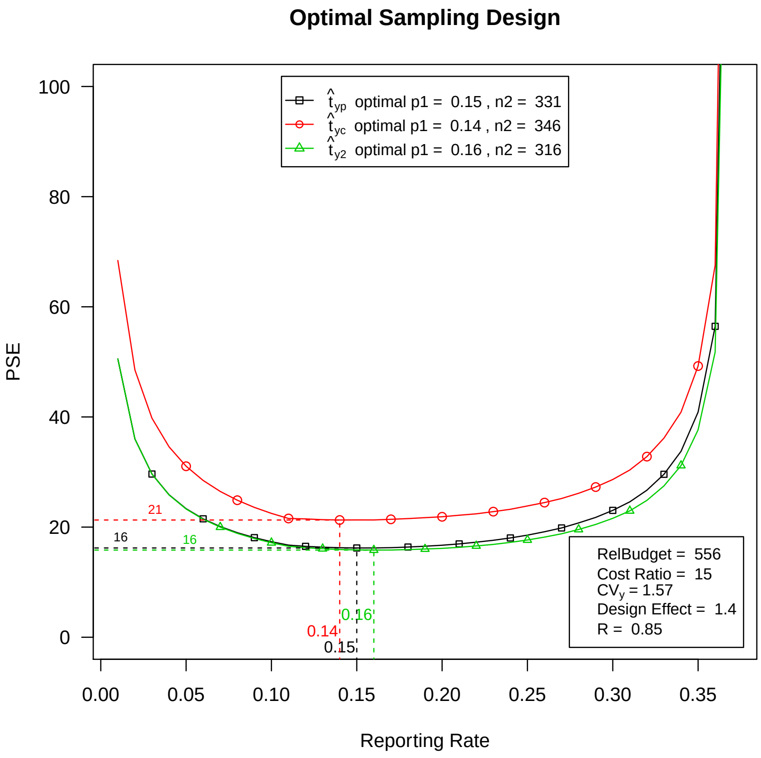


Figure 3.5: Optimal pilot sampling design for total recreational Red Snapper catch by Charter boat.

3.4. Discussion

The approach implemented in the R package *OptimalFisheryDesign* allows the user to calculate the sample characteristics, investigate the effect of the two sample sources, understand the trade-offs between the two sample sources, compare the performance of all the three estimators, and determine their optimal sampling strategies. However, the sample designer should be cautious when designing the survey based the package for several reasons.

First, the sample designer may not be able to provide an accurate cost ratio, which may compromise the conclusions from the optimal designs provided by the package. For example, the cost of collecting the self reports involves many aspects such as the cost of setting up and maintaining the electronic reporting system. As a result, it may be hard to estimate the average cost for increasing the reporting rate by one percent compared to the cost for recruiting one more dockside intercept sample unit. As a solution, we recommend that the researcher examine a range of values for the cost ratio when determine how to best distribute their budget.

Second, the sample designer should notice that the package assumes a constant cost ratio for different reporting rates, which may not be true in practice. For example, when the current reporting rate is low, it may be feasible to increase the reporting rate by putting more ads or setting a regulation that requires the anglers to report. On the other hand, when the current reporting rate is high, it may be difficult to increase the reporting rate further with any new policy. This is because the anglers who don't report in this situation are more likely to be the ones who are reluctant to do so. However, it is still reasonable to assume the cost ratio is relatively stable within a certain range. For example, the cost ratio may not change dramatically when the reporting rate increases from 10% to 15%.

Last, the sample designer should be aware that the package does not consider the effect of non-sampling errors, which may bias the conclusions. For example, the estimation approach requires matching of units in the two samples, an operation that can prove difficult in practice. The quality and quantity of the matched units will affect the estimation precision of $\bar{y}_1$, CV_{1y}, R_{1,yy^*}, which are the inputs of different functions in the package. Another challenge of the design is how to control the independence of the two sample responses; i.e., to assure that reporting is not influenced by the trip being included in the intercept sample. All three estimators require this assumption for validity. Detailed discussions and recommendations about such issues can be found in Stokes et al. (2019b).

CHAPTER 4

Alternative Estimation of Total Approaches Based on Non-probability Sampling

In this chapter, we investigate the viability of using the self-reported sample (non-probability sample) to augment the size of the dockside intercept sample (probability sample) for estimating the total fish catch. However, this approach only works for highly recognizable species since we must assume that the self reports are accurate. This led us to the non-probability sampling literature.

There has been much recent research in the sampling literature on how to make use of non-probability sample data in a more principled way. This is due to the increased availability of inexpensive, easily accessible, and large sets of data from various data collection operations, such as internet samples, electronic device data (voluntarily or involuntarily produced), and operational data of various forms. We briefly review them in the next section.

4.1. Literature Review of Non-probability Sample Estimation Methods

There are four main approaches to make estimates from non-probability samples: design based, model based, doubly robust and calibration. All the approaches attempt to improve estimation by using auxiliary information and rely on either a reference sample or the parameters of the population. The reference sample is usually a high quality probability sample from the population, but may not contain the outcome variable. Here we briefly review the four approaches.

4.1.1. Design Based Approach

The design based approach refers to creation of pseudo-inclusion probabilities for the non-probability sample, which is usually done by propensity score adjustment (PSA) (Lee, 2006; Lee and Valliant, 2009; Schonlau et al., 2009; Valliant and Dever, 2011). The propensity score model was originally developed for observational studies to reduce the bias due to confounding variables (Rosenbaum and Rubin, 1983). In the context of non-probability sampling, the PSA method is carried out by combining the non-probability sample with a probability sample, and estimating the selection probability for the non-probability sample. The estimated probability is then used to create pseudo-weights for the units in the non-probability sample. This allows a Horvitz-Thompson type estimator to be constructed, which is known as the inverse probability weighted (IPW) estimator. To build the propensity score model, the non-probability sample and probability sample are required to share a set of covariates. Although sometimes not explicitly stated in the current studies, the PSA method assumes no overlap between the two sample sources, otherwise it will be ambiguous to determine the membership of the sample units. This assumption is reasonable when the two samples are from a very large population, since then the chance for a sample unit to be included in both samples is negligible. However, this assumption could be violated in practice when the population size is relatively small, as in our case.

In order for the PSA method to be effective at eliminating selection bias, the strong ignorability assumption must be met. This assumption requires that the selection mechanism for the non-probability sample be independent of the outcome variable, either unconditionally or conditionally on other covariates that are observable. When the selection probabilities are accurately estimated, the resulting IPW estimator is asymptotically unbiased. However, this assumption can be violated in two ways: 1) when the inclusion probability for the non-probability sample depends on the response variable, which is the case of not missing at random (NMAR); 2) when not every sample unit in the population has a positive probability to be included in the non-probability sample.

Unfortunately, the two situations are commonly encountered in the context of non-probability data. For example, people who do recreational fishing tend to self-report their fish catch if they have a harvest, and not otherwise. So unless there are other available variables that predict the trips with no catch, the PSA method will fail to remove the selection bias from the self-reported data. The second situation can be illustrated by the fact that not every fisherman in the recreational fishery population has a self-reporting device (such as a cell phone)(Liu et al., 2017), so they have zero chance to be included in the non-probability sample. On the other hand, if the response variable for those that are included is similar to those with no probability of inclusion, the procedure could still perform well.

4.1.2. Model Based Approach

The model based approach aims to predict the outcome variable from the auxiliary information and then project the non-probability sample to the target population (Valliant, 2019). The generalized regression model (GREG) is commonly used for such a prediction purpose. For estimating the population total, this is done by modeling the relationship between the outcome variable and covariates based on the non-probability sample, estimating the outcome variable for the non-sampled units, and then summing the predictions for the non-sampled units and the outcomes from the non-probability sample. To construct the estimator, the population totals of the covariates are required, which can be obtained from census data or estimated from a reference sample from other sources that contain such information.

Compared to the design based approach where individual level information is required for the reference sample, the model based approach only requires the population level summary statistics for some key variables. Another distinction between the two approaches is that the model based approach assumes the randomness is from the model instead of the process of generating the non-probability sample. As a result, if the relationship between the outcome variable and the covariates was correctly modeled, the estimator from the model based approach is asymptotically unbiased. This advantage makes the model based approach promising when the data to be analyzed has rich auxiliary information. However, the model based approach relies on the assumption that the model fitted by the non-probability sample also fits the non-sampled units. If the selection mechanism for the non-probability sample is NMAR, then the models for the non-probability sample and non-sampled units might be different.

4.1.3. Doubly Robust Estimation

As a combination of the design based approach and model based approach, doubly robust estimation has also been proposed in the context of non-probability sampling (Chen et al., 2018). As the name suggests, doubly robust estimator is asymptotically unbiased if either the pseudo-inclusion probability, the generalized regression model or both are correctly specified. Valliant (2019) demonstrated through simulation studies that the doubly robust estimator is generally the least biased with smallest mean squared error (MSE) compared to the estimators from PSA or GREG.

4.1.4. Calibration

The calibration approach aims to adjust the auxiliary information in the non-probability samples to the probability sample such that the weighted distribution for covariates in the non-probability sample is similar to that of the population. General calibration methods involve post-stratification (Bethlehem, 2010), raking and generalized regression weighting

(Deville and Särndal, 1992). Valliant and Dever (2011) demonstrated through simulation studies the potential for bias reduction by calibrating the non-probability sample, and its performance is comparable to PSA. They observed that, as with PSA, when the inclusion probability for the non-probability sample is associated with the outcome variable, the calibration adjustment will fail.

4.1.5. Other Approaches

There are other approaches to make inference from the non-probability sample. Rivers (2007) proposed the idea of nearest neighbor matching, which selects sample units from the non-probability sample that have similar auxiliary information to the sample units in the reference sample. The selected non-probability sample units are then expected to mimic the characteristics of the reference sample. Liu et al. (2017) took a different approach of using a non-probability sample to estimate total recreational red snapper catch in Texas. As described previously, they proposed the ratio estimators that augments data from a probability sample with data from an overlapping non-probability sample. The non-probability sample is treated as auxiliary variables in the ratio estimators. Thus, the estimators do not require either weighting adjustment for the non-probability sample or the representativeness of the non-probability sample. However, their approach does require matching of units in the two samples, an operation that can prove difficult in practice.

The effectiveness of the aforementioned methods depends on the quality of auxiliary information, such as how well it can explain the selection mechanism or the outcome variable. As summarized by Baker et al. (2013), there are still many obstacles that need to be overcome while dealing with non-probability samples: (i) unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling; (ii) making inferences for any non-probability sample require some reliance on modeling assumptions; (iii) if non-probability samples are to gain wider acceptance among survey researchers, there must be an accompanying set of measures for evaluating their quality.

Our application differs in several ways from the typical scenarios considered in non-probability sampling. Most of the current approaches are conducted for the situation that the outcome variable is available for the non-probability sample, where the reference sample only contains the auxiliary information. In our application, the outcome variable is also available in the reference sample. We can take advantage of this extra piece of information by using the outcome variable in estimation and treating its value as a benchmark. Our reference sample is relatively small, which makes the model based approach less efficient, as we will have to build the prediction model on the reference sample and predict the outcome variable for the non-probability sample. Another distinction of our application is the non-negligible overlap between the non-probability sample and the reference sample, whose inclusion probability for the non-probability sample is also need to be estimated.

4.2. Pilot Study

In our application, we focus on the design based approach and aim to estimate the pseudo-inclusion probability for the non-probability sample, so that the two samples can be treated together as an augmented probability sample. The design based approach makes use of the additional information, which can be in the form of a probability sample from the same population or even census level information. Because our self-reported sample is so large, the question arises as to whether these methods could be beneficial in our application.

To gain an idea of how much improvement over the current estimators is possible if we could "convert" our non-probability sample to a probability one, we conducted a simple pilot study. Our goal was to determine whether or not the potential advantage, making the most optimistic assumptions, is large enough to suggest this would be a fruitful approach to pursue. The simplified scenario we examined included the following assumptions: (a)

$y^* = y$ (perfect reporting), (b) the self-reported sample (S_1) behaves like a simple random sample, and (c) our pseudo selection probability (n_1/N) is accurately estimated. We also suppose for simplicity that our probability sample is a simple random sample. In this case, the combined sample $S = S_1 \cup S_2$ is also a simple random sample from the population with sample size $n_S = n_1 + n_2 - m$, where m is the overlap sample size as we defined in Subsection 2.2.2. Then a reasonable estimator of total $\hat{t}_{SRS}$ has the same form of $\hat{t}_{yp\_SRS}$:

$$\hat{t}_{SRS} = \hat{N}\bar{y}_S, \tag{4.1}$$

where $\hat{N} = \frac{n_1 n_2}{m}$ and $\bar{y}_S = \sum_{i\in S} y_i/n_S$, which is the sample average of the $y's$ from the combined sample $S = S_1 \cup S_2$. The variance of $\hat{t}_{SRS}$ (see Appendix B) can be approximated by:

$$V(\hat{t}_{SRS}) \approx \frac{N^2}{E(n_S)}\{S_y^2(1 - \frac{E(n_S)}{N}) + \bar{y}^2 \frac{E(n_S)}{n_2}\frac{(1-p_1)}{p_1}(1 - \frac{n_2}{N})\}, \tag{4.2}$$

where $E(n_S) = n_1 + n_2 - \frac{n_1 n_2}{N}$. Under this simplified scenario, we have $\bar{y} =\bar{y}_1^* = \bar{y}_1,\ CV_y = CV_{1y} = CV_{1y^*}$ and $R_{1,yy^*} = 1$. Thus the variance of $\hat{t}_{y2}$ becomes easy to compare to that of $\hat{t}_{SRS}$.

To compare $V(\hat{t}_{y2})$ with $V(\hat{t}_{SRS})$, we set $N = 15000$, $n_2 = 400$ and $\bar{y} = 10$. Figure 2.1 displays the ratios of $V(\hat{t}_{y2})/V(\hat{t}_{SRS})$ as functions of the reporting rate (p_1) for a range of $CV_y :\ 1,\ 2,\ 3,\ 4$. The ratios are greater than 1 for all conditions, which shows that $\hat{t}_{SRS}$ is less variable compared to $\hat{t}_{y2}$. This meets our expectation, since the reported data are fully used as sample data rather than just as auxiliary data as $\hat{t}_{y2}$ does. As the reporting rate (p_1) increases, the variance ratios also increase, which indicates that $\hat{t}_{SRS}$ will have a greater advantage compared to $\hat{t}_{y2}$. Figure 2.1 also illustrates that $\hat{t}_{SRS}$ has more advantage than $\hat{t}_{y2}$ when the population has a larger CV_y, for the same reporting rate. This information is valuable as it allows us to consider using different estimators for different fish species. For example, $\hat{t}_{y2}$ may be preferable for estimating catch for a

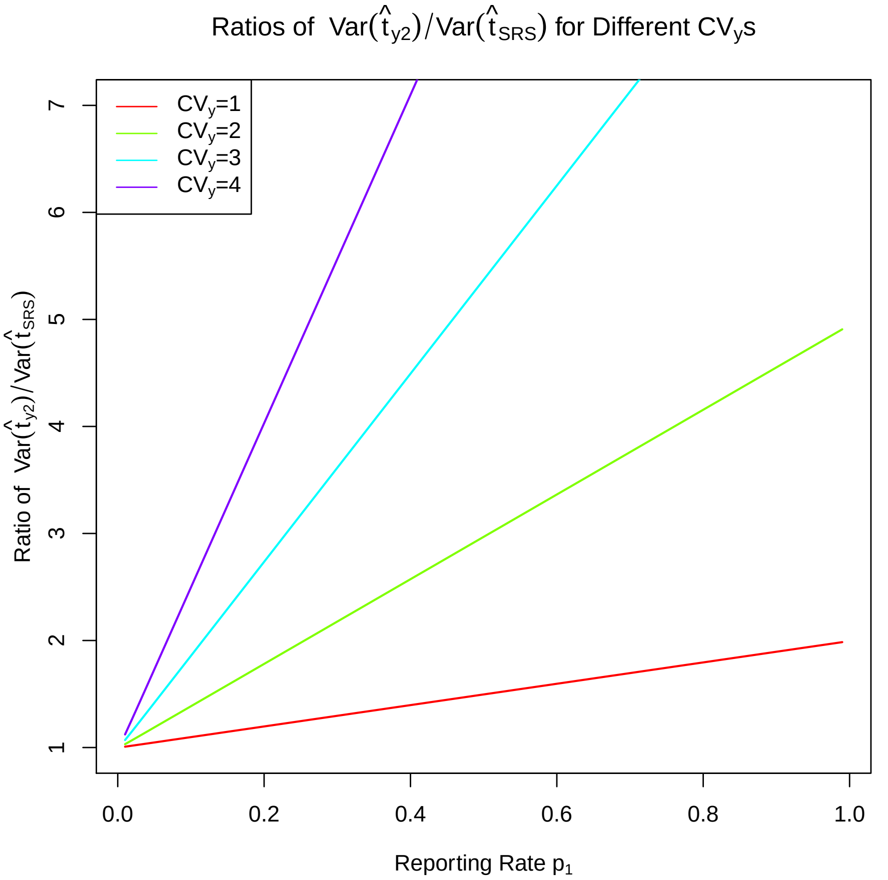


Figure 4.1: Effect of the reporting rate (p_1) on the variance ratio $V(\hat{t}_{y2})/V(\hat{t}_{SRS})$ for different CV_ys: 1, 2, 3, 4.

fish species with small CV_y, like Red Snapper. However, there are many other valuable fish species, like Spanish Mackerel and Vermilion Snapper, and especially rare species, whose CV_y are large.

Of course in a real application, the selection probability for the non-probability sample will have to be estimated using a method such as PSA, which undoubtedly will decrease the advantage. In addition, if the available auxiliary information does not completely explain the selection mechanism so that bias is introduced, the new estimator will be further disadvantaged. However, this exercise suggests that the most promising place to look for possible improvement from this approach is in estimating catch for fish species with large

CV_y, which are the ones where the current method is sometimes inadequately precise.

4.3. Propensity Score Adjustment (PSA)

Let

$$\alpha_i = P\left(i \in S_2 | \mathbf{X}_i\right), \ i \in 1, 2, ..., N \tag{4.3}$$

denote the inclusion probability for the i^{th} sample unit in the probability sample S_2, and its weight is the reciprocal of the inclusion probability:

$$w_i = 1/\alpha_i = 1/P\left(i \in S_2 | \mathbf{X}_i\right), \ i \in 1, 2, ..., N. \tag{4.4}$$

This weight is known from the probability sampling design if the analyst is the sample designer. If the analyst is only a secondary data user, he will still know the weight if it is included on the data file. In our application, this weight is known.

We denote the selection probability for the i^{th} sample unit in the non-probability sample S_1 by:

$$\beta_i = P(i \in S_1 | \mathbf{X}_i), \ i \in 1, 2, ..., N. \tag{4.5}$$

Then the selection probability for the unit i in the combined sample $S = S_1 \cup S_2$ can be expressed as:

$$q_i = P\left(i \in S | \mathbf{X}_i\right) = \alpha_i + \beta_i - \alpha_i \beta_i, \ i \in 1, 2, ..., N. \tag{4.6}$$

We also denote the conditional probability for the i^{th} unit in the non-probability sample (S_1) given that it is in the combined sample (S) by:

$$\gamma_i = P(D_i = 1 | i \in S, \mathbf{X}_i) = \frac{\beta_i}{\alpha_i + \beta_i - \alpha_i \beta_i}, \ i \in 1, 2, ..., N, \tag{4.7}$$

where $D_i = 1$ if $i \in S_1$ and $D_i = 0$ otherwise.

To create the pseudo-weight for sample unit i, the inclusion probabilities β_i, q_i and γ_i must be estimated. First, the conditional probability γ_i can be estimated by a propensity score model. We denote this design-based estimator by $\hat{\gamma}_i$. By solving equation (4.7), we can estimate the inclusion probability β_i by:

$$\hat{\beta}_i = \frac{\alpha_i \hat{\gamma}_i}{1 - \hat{\gamma}_i + \alpha_i \hat{\gamma}_i},\ i \in 1, 2, ..., N. \tag{4.8}$$

Thus, the probability that sample unit i is included in the combined sample can be estimated by:

$$\hat{q}_i = \alpha_i + \hat{\beta}_i - \alpha_i \hat{\beta}_i = \frac{\alpha_i}{1 - \hat{\gamma}_i + \alpha_i \hat{\gamma}_i},\ i \in 1, 2, ..., N. \tag{4.9}$$

Calculation of $\hat{\beta}_i$ and $\hat{q}_i$ from equations (4.8) and (4.9) require that α_i is also known or estimable for unit i from the non-probability sample. Even though this quantity is usually unknown, it could be determined in our application. This is because the selection probability α_i is controlled by the dock location and return time of the trip, which are contained in the self reports. When α_i is unknown for the i^{th} unit in the non-probability sample, Robbins et al. (2015) suggested to assign the same inclusion probability to every non-probability sample units, which is the average of the inclusion probabilities from the probability sample units. Other approaches to estimate this quantity can be found in Elliott et al. (2017).

Equipped with the estimated inclusion probability from equation (4.9), the non-probability sample can be weighted to represent the population either by itself or combined with the probability sample, which leads to two weighting schemes: joint weighting and disjoint weighting. Robbins et al. (2015) proposed the two approaches for estimating the population mean, and they found that joint weighting always provide smaller estimation variance while disjoint weighting can be used to detect whether the two samples are well integrated by comparing their estimates. We adapt the two weighting schemes into our application for estimating the population total with unknown population size. We now describe our two proposed estimation approaches.

4.3.1. Joint Weighting

In joint weighting, we construct a single estimate of total based on the combined sample: $S = S_1 \cup S_2$. Thus, the inclusion probability for the i^{th} unit in the combined sample S is estimated by $\hat{q}_i$ in equation (4.9), and its estimated pseudo-weight is the reciprocal of $\hat{q}_i$:

$$\hat{w}_{i,joint} = 1/\hat{q}_i, \ i \in S = S_1 \cup S_2.$$

Consequently, we can construct the Horvitz-Thompson type estimator:

$$\hat{t}_{y,joint\_HT} = \sum_{i \in S} \hat{w}_{i,joint} y_i. \tag{4.10}$$

In our application, however, the estimator in equation (4.10) will be biased downward due to undercoverage in the dockside sample, as discussed previously. Thus the joint weighting estimator of total we propose takes the form of $\hat{t}_{yp}$ and has the expression:

$$\hat{t}_{y,joint} = \hat{N}\widehat{\bar{y}} = \frac{n_1}{\hat{p}_1}\widehat{\bar{y}} = \frac{n_1}{\sum_{i \in S_2} w_i r_i / \sum_{i \in S_2} w_i} \frac{\sum_{i \in S} \hat{w}_{i,joint} y_i}{\sum_{i \in S} \hat{w}_{i,joint}}, \tag{4.11}$$

where $\hat{N}$ is the estimator of the population total, $\widehat{\bar{y}}$ is the estimator of the mean catch and r_i is the reporting indicator as we defined in Subsection 1.4.2. Thus, the role of the non-probability sample is to improve estimation for the mean catch, but not population size.

4.3.2. Disjoint Weighting

In disjoint weighting, each sample is weighted to be individually representative of the population. The disjoint weighting estimator of total is then a weighted average of the estimators of total from both samples. For the non-probability sample (S_1), the estimated pseudo-inclusion probability for the i^{th} unit is $\hat{\beta}_i = \frac{\hat{\alpha}_i\hat{\gamma}_i}{1-\hat{\gamma}_i+\hat{\alpha}_i\hat{\gamma}_i}$, $i \in S_1$, and its weight can be expressed as $\hat{w}_{i,disjoint} = 1/\hat{\beta}_i, i \in S_1$. If undercoverage was not a problem, the natural estimator of total would be a weighted average of the two Horvitz-Thompson type estimators:

$$\hat{t}_{y,disjoint\_HT} = \theta \sum_{i\in S_1} \hat{w}_{i,disjoint} y_i + (1-\theta) \sum_{i \in S_2} w_i y_i. \tag{4.12}$$

In our application, however, we must protect against undercoverage in the probability sample, and thus propose the estimator of total $\hat{t}_{y,disjoint}$, which takes the form of $\hat{t}_{yp}$ and has the expression:

$$\hat{t}_{y,disjoint} = \hat{N}\widehat{\bar{y}} = \frac{n_1}{\hat{p}_1}(\theta \frac{\sum_{i\in S_1} \hat{w}_{i,disjoint} y_i}{\sum_{i\in S_1} \hat{w}_{i,disjoint}} + (1-\theta)\frac{\sum_{i\in S_2} w_i y_i}{\sum_{i\in S_2} w_i}). \tag{4.13}$$

As suggested by Robbins et al. (2015), the parameter θ is chosen to minimize the Kish approximation of the estimator's design effect and has the form:

$$\theta = \frac{\left(\sum_{i\in S_2} \alpha_i^{-1}\right)\left(\sum_{i\in S_1} \hat{\beta}_i^{-2}\right)}{\left(\sum_{i\in S_2} \alpha_i^{-1}\right)\left(\sum_{i\in S_1} \hat{\beta}_i^{-2}\right) + \left(\sum_{i\in S_2} \alpha_i^{-2}\right)\left(\sum_{i\in S_1} \hat{\beta}_i^{-1}\right)}. \tag{4.14}$$

4.4. Adaptive Propensity Score Adjustment (APSA)

The propensity score adjustment (PSA) method was developed to help reduce the selection bias of a non-probability sample (e.g., Valliant and Dever, 2011; Lee and Valliant, 2009). However, this method will not remove the bias if the strong ignorability assumption on which the propensity score model relies is violated. In our application, for example, if the self-reported trips are only from the boats with more than 15 anglers, then the estimated inclusion probability for the self-reported sample is unreliable for the boats with less than 15 anglers. This is either because the trips with less than 15 anglers have zero probability to self report or the self-reported sample does not allow that probability to be estimated even if it is positive, as there are no such sample cases in the self-reported sample. However, if we have some evidence that the trips with less than 15 anglers can be represented by part of the trips with more than 15 anglers, the strong ignorability assumption still holds.

Here we propose a new approach called the adaptive propensity score adjustment (APSA) method. Our method is based on the current PSA method, but adds a step that provides an indicator of model failure. The APSA method is also designed to reduce the selection bias by using only part of the non-probability sample for PSA, as we will describe later. When the outcome variable y is available for both the probability and non-probability samples, the APSA method uses the fact that when the selection probability is correctly estimated by the propensity score model, sample units with similar propensity scores should have similar outcome values. We will first justify this argument and then describe the new method.

4.4.1. Comparison between Causal Inference and Non-probability sampling

In the context of non-probability sampling, the propensity score $\gamma_i = P(D_i = 1|i \in S, X_i)$ is the conditional probability that unit i is in the non-probability sample (S_1) given that it is in the combined sample (S) and has auxiliary information X_i. This setup fits the non-probability sample into the causal inference framework: the non-probability sample can be treated as the treatment group and the probability sample is treated as the control group.

As in causal inference, we assume two potential outcomes for the sample units: y_0 if the sample unit i is in the probability sample and y_1 if it is in the non-probability sample. The strong ignorability assumption for the self-reporting mechanism in the non-probability sampling context can be stated as: $(y_1, y_0) \perp D_i \,|\, i \in S, X_i$ and $0 < \gamma < 1$. In plain words, this means that 1) all possible confounding variables which affect both the non-probability sample selection mechanism and the outcome variable are measured in X_i, 2) there is a positive probability for every unit in the population to be selected into the non-probability sample.

In causal inference, the quantity to be estimated is the average treatment effect, defined as $E(y_1) - E(y_0)$. However, in non-probability sampling, whether the sample unit is selected into the probability sample or non-probability sample does not affect the outcome value and we always have: $y_0 = y_1$ and $E(y_1 - y_0)$ = 0. In practice, we can only observe either y_0 or y_1 if there is no overlap between the probability sample and the non-probability sample. Under the strong ignorability assumption, the expected difference between the outcomes from the probability sample and the non-probability sample is a function of the estimated propensity score γ_i, and can be expressed as:

$$E(y_1|\gamma_i, D_i = 1) - E(y_0|\gamma_i, D_i = 0) = E(y_1 - y_0|\gamma_i). \tag{4.15}$$

As the outcome for a sample unit does not change regardless of its membership, the last term in equation (4.15) equals 0. The above argument established the following theorem in the context of non-probability sampling, which corresponds to Theorem 4 established by Rosenbaum and Rubin (1983) in the context of causal inference.

Theorem 1 *Suppose the selection mechanism for the non-probability sample is strongly ignorable and γ_i is the propensity score. Then the expected difference in the observed outcome between the probability sample and non-probability sample at γ_i is equal to 0, that is,*

$$E(y_1|\gamma_i, D_i = 1) - E(y_0|\gamma_i, D_i = 0) = 0.$$

Even though our purpose is not comparing the difference between the probability sample and the non-probability sample outcomes, the theorem provides a way to check whether the strong ignorability assumption is violated. If a non-zero difference exists between the means from the probability sample and non-probability sample with the same propensity score, it indicates a model failure and thus the propensity score adjustment is unreliable. This comparison can be conducted by post-stratification based on the propensity scores (Rosenbaum and Rubin, 1984), and will be described in the next section.

4.4.2. Adaptive Propensity Score Adjustment (APSA) Algorithm

The first step in the proposed algorithm is to post-stratify the probability sample and non-probability sample into subgroups such that the propensity scores within each subgroup are similar. If the strong ignorability assumption holds, the means of the response y from the two samples within each subgroup are expected to be similar. If the response from the two samples are different within a certain subgroup, the corresponding non-probability sample units will be dropped. This is because their selection mechanism cannot be explained by the current model, and thus makes the selection bias non-adjustable.

Here we describe the APSA algorithm that performs the above procedure:

Step 1: Calculate the propensity scores for every unit of the full sample $S = S_1 \cup S_2$,

Step 2: Sort the estimated propensity scores from smallest to largest and form the sample into 10 subgroups by decile points,

Step 3: Within each subgroup, conduct hypothesis testing to compare the means of the outcome variable between the two samples,

Step 4: Identify the subgroups with significant p-values, discard units from the non-probability sample but keep the units from the probability sample for such subgroups,

Step 5: Re-calculate propensity scores for the retained sample and re-conduct the PSA method.

From APSA, the new combined sample contains the full probability sample and the non-discarded part of the non-probability sample. Based on the new combined sample, we re-calculate the estimators from the joint weighting and the disjoint weighting approaches, and denote them by $\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$. If the probability sample did not have under-coverage, we could also compute the Horvitz-Thompson type estimators based on the new combined sample using joint weighting or disjoint weighting, which are denoted by $\hat{t}_{y,joint\_HT\_adp}$ and $\hat{t}_{y,disjoint\_HT\_adp}$.

We also note that the non-discarded part of the non-probability sample cannot be adjusted to represent the population on its own, which makes $\hat{t}_{y,disjoint\_adp}$ and $\hat{t}_{y,disjoint\_HT\_adp}$ unreliable. So we should not use the two estimators if we discard some of the non-probability samples.

4.5. Variance Estimation

Now, we describe two replication based methods to estimate the variance of the proposed estimators, and then evaluate them by simulation studies based on data similar to that of the motivating example.

Taylor series linearization is often used for variance estimation in survey practice, and it has been applied in several non-probability sampling studies. However, as demonstrated by simulation studies, the method underestimates the actual variance of estimators of means from non-probability samples (e.g. Lee and Valliant, 2009, Robbins et al., 2015). This is because the method treats the estimated propensity score as the actual selection probability, and thus ignores its estimation variance. Alternative approaches to variance estimation can be conducted by the jackknife and bootstrap methods (e.g. Kim and Wang, 2019, Valliant, 2019). They are resampling techniques that may be able to reflect the added variation from the estimated propensity scores.

For the jackknife method, Section 4.4 of Wolter (2007) (page 169) states that no theory actually justifies the jackknife method for nonlinear estimators for general complex sample designs. However, it has been shown to work well empirically, and there are theoretical results for some simple complexities, such as strata. To apply the empirically based jackknife method when there is a complex design, we leave out K PSUs at a time within a stratum and redistribute their weight across other units in that stratum, even if the PSUs are unequally weighted. We apply this method in our application to examine its performance.

We apply this empirically based jackknife method to estimate the variance of the PSA estimators $\hat{t}_{y,\ joint}$ and $\hat{t}_{y,\ disjoint}$. In our application, the probability sample contains design information about PSUs and strata (Weekday/Weekend), but the non-probability sample only contains information about strata. Therefore, we adapt the jackknife method by treating each trip from the non-probability sample as a PSU. Then for each jackknife replicate, we leave out one PSU from the probability sample and $K-1$ PSUs from the non-probability sample from the same stratum. The value of K is determined by the number of PSUs in the probability sample.

However, there is a problem when applying a jackknife for variance estimation to the APSA estimators $\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$. An intuitive justification for the jackknife as a method of variance estimation is that the leave-K-out pseudo-values are independent and identically distributed. When a different number of observations are dropped from the sample in different replicates, they are based on different sample sizes. Thus they can not be justified as even approximately identically distributed. Therefore, we adapt the jackknife method by conditioning on the retained sample from APSA to estimate the variance of $\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$. The retained sample from each replicate has a constant sample size, which makes the intuitive notion of iid pseudo-values more plausible. This is done by the conditional decomposition:

$$\begin{aligned} Var(\hat{t}_{y,joint\_adp}/\hat{t}_{y,disjoint\_adp}) = & E[Var(\hat{t}_{y,joint\_adp}/\hat{t}_{y,disjoint\_adp}|retained\ sample)] \\ & + Var[E(\hat{t}_{y,joint\_adp}/\hat{t}_{y,disjoint\_adp}|retained\ sample)]. \end{aligned} \tag{4.16}$$

The first term on the right hand side of equation (4.16) is estimated by applying the jackknife method on the retained sample while treating the APSA estimators as PSA estimators, so the dropping mechanism from the APSA method is not applied here. Since it is almost impossible to enumerate all possible retained samples from the APSA method, the second term on the right hand side of equation (4.16) is approximated by $Var(E(\hat{t}_{y,joint\_adp}/\hat{t}_{y,disjoint\_adp}|number\ of\ retained\ subgroups))$, which is estimated from the jackknife replicates of the original sample.

For the bootstrap method, we apply the standard procedure as described in Wolter (2007) to estimate the variance of the PSA estimators $\hat{t}_{y,\ joint}$ and $\hat{t}_{y,\ disjoint}$. This is done by resampling the PSUs from each stratum for each replicate. For the APSA estimators $\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$, we apply the bootstrap method based on the conditional decomposition (equation 4.16), which is the same way as we apply the jackknife method.

4.6. Simulation Studies

4.6.1. Simulation Settings

So far, we have proposed 4 estimators that use the non-probability sample, for cases where under-coverage in the probability sample is a concern as in our application. They are: $\hat{t}_{y,\ joint}$, $\hat{t}_{y,\ disjoint}$,$\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$, all of which are alternatives to the current ratio estimators. The four estimators differ in how they weight the units from the two samples, including whether they give the unit a non-zero weight at all. To assess the efficiency of the different weighting strategies, we conducted two simulation studies to evaluate the four estimators in situations where the strong ignorability assumption is and is not met. The first simulation study aimed to study the performance of the four estimators when the probability sample is a SRS. The second simulation study aimed to study the performance of the four estimators when the probability sample is drawn according to a complex design. We used the stratified cluster design with PSUs selected randomly from each stratum in this simulation. This design mimicked the actual design that was used by the Access Point Angler Intercept Survey (APAIS) to collect the dockside intercept sample. In the simulation studies, we used the ratio estimator $\hat{t}_{y2}$ as our benchmark for comparison. This is because $\hat{t}_{y2}$ was shown to have good performance under a wider range of scenarios than the other two ratio estimators (Liu et al., 2017).

To maintain the relationship between the catch and auxiliary information for the trips, a pseudo-population of recreational fishing trips was formed by using all the self reports from a NMFS experimental electronic logbook study (ELB). This study was conducted to estimate the catch of all Gulf of Mexico fish species caught by charter boat from 2016 - 2017. The self reported sample contained 15771 trip records. We treated this data as if it was the true population. In addition to the catch counts by species, the self reports contained several other variables describing the trip, which are listed in Table 4.1. The variable of interest y is defined as the sum of the fish caught ($Kept$) and discarded dead ($ReleasedDead$) for all species for each trip. We refer to this variable as catch. To avoid extreme values in self-reporting, we truncated every numerical variable by the 1.5 IQR rule except the outcome variable y.

In this pseudo-population, the primary sampling unit (PSU) was defined as a combination of location ($County$), Wave ($Wave$), return time ($Shift$) and time period ($Weekend$). If a PSU contained more than 100 trips, it was randomly segmented into several PSUs to make sure all PSUs had size less than 100. This definition aimed to mimic the actual PSUs collected by the Access Point Angler Intercept Survey (APAIS) for the dockside intercept sample. In this population, there were 956 PSUs and the average number of trips per PSU was 16. The strata were defined as whether the trip was made on a weekday or weekend. There were 495 (52%) PSUs from a weekday and 461 (48%) PSUs from a weekend.

To compare the performance of $\hat{t}_{y,joint}$, $\hat{t}_{y,disjoint}$, $\hat{t}_{y,joint\_adp}$,$\hat{t}_{y,disjoint\_adp}$ with $\hat{t}_{y2}$, both simulation studies followed a $4 \times 4 \times 4$ factorial design. The three factors are: 1) probability sample size, which is the dockside intercept sample size in our application. 2) non-probability sample size, provided as reporting rate of the self-reported sample in our application $(p_1 = n_1/N)$; 3) non-probability sample selection mechanism, which is the self reporting mechanism in our application.

Table 4.1: Description of variables in the pseudo-population for simulation.

Variable Name	Type	Model Inclusion	Description
Trip ID	Cat.	No	Identification number of each trip
Kept	Cont.	Yes	Fish caught by species
Release	Cont.	Yes	Alive fish released by species
ReleasedDead	Cont.	Yes	Dead fish released by species
CaptainName	Cat.	No	Captain's name of the boat
Latitude	Cont.	No	Latitude when self-reported
Longitude	Cont.	No	Longitude when self-reported
NbPassengers	Cont.	Yes	Number of passengers on the boat
NbAnglers	Cont.	Yes	Number of anglers on the boat
NbCrew	Cont.	Yes	Number of Crew on the boat
DepthPrimary	Cont.	Yes	Depth of the sea when fishing
Hours	Cont.	Yes	Fishing duration
Shift	Cat.	No	Return time: 1 (00:00 AM - 8:00 AM) , 2 (8:00 AM - 4:00 PM), 3 (4:00PM - 12:00 PM)
Weekend	Cat.	Yes	Weekend: 1, Weekday:0
State	Cat.	Yes	Home state of the boat: AL, FL, LA, MS, TX
County	Cat.	No	Home county of the boat
Wave	Cat.	No	Fishing waves: 1 - 6
Name	Cat.	No	Name of the boat

For the first factor, the probability sample sizes (n_2) were set to 200, 400, 600 and 800 for the SRS design. For the complex design, the probability sample sizes (n_{PSU}) were determined by the number of PSUs, which were set to 30, 40, 50 and 60. For the second factor, the non-probability sample sizes were set to 3154, 4731, 6308 and 7885, which corresponded to the four reporting rates: $p_1 = 0.2,\ 0.3,\ 0.4,\ 0.5$. For each combination of the probability sample and non-probability sample sizes, 4 different non-probability sample selection mechanisms were examined from the missing data perspective: Missing Complete at Random (MCAR), Missing at Random (MAR) and two that are Not Missing at Random (NMAR), one of which we denoted as the Large Catch case. Now we describe the 4 selection mechanisms.

1) Missing Completely at Random (MCAR)

This scenario was intended to examine the performance of the proposed estimators when the non-probability sample is free of selection bias. It was conducted by drawing the non-probability sample as a simple random sample from the population. In this situation, the selection mechanism of the non-probability sample is independent of the outcome variable unconditionally upon the auxiliary information X, which is analogous to the notion of data that is missing completely at random.

2) Missing at Random (MAR)

This scenario was intended to examine the performance of the proposed estimators when the selection mechanism of the non-probability sample is independent of the outcome variable conditionally upon the auxiliary information X, which corresponds to the notion of data that is missing at random (MAR). This scenario was conducted by generating the selection probability of the non-probability sample sample from a pre-specified model:

$$log(\frac{\beta_i}{1-\beta_i}) = 0.3\times NbPassengers + 0.2\times Release + 0.1\times Hours - 0.2\times Weekend, \;\; i = 1, 2...N. \tag{4.17}$$

In this model, β_i is the probability for the i^{th} trip to be included in the non-probability sample. The correlation between the selection probability and outcome variable y from this model is 0.42. For each reporting rate, the generated selection probabilities were multiplicatively adjusted to sum to the expected non-probability sample size.

3) Large Catch Case

This scenario was intended to examine the bias reduction ability of the proposed estimators when the non-probability sample contains large selection bias. To simulate this situation, we partitioned the population into two strata: the first stratum contained the trips with the largest fish catch of the population, and the size of the stratum was set to be 30%

of the non-probability sample size. The second stratum was the rest of the population. For the non-probability sample of sizes 3154, 4731, 6308 and 7885, the sizes of the first stratum were set to 946, 1419, 1892 and 2365, respectively. The non-probability sample was drawn as a stratified sample from the population, and it contained all sample units from the first stratum and a simple random sample from the second stratum. In this scenario, even though the inclusion mechanism depends on the response variable y, it allows every unit in the population to have a positive inclusion probability.

4) Not Missing at Random (NMAR)

This scenario was intended to examine the performance of the proposed estimators when the non-probability sample has under-coverage and its selection mechanism depends on the outcome variable y. To simulate this situation, we partitioned the population into two strata based on whether or not the number of fish catch was greater than 7. The small-catch stratum contained 2602 trips and the large-catch stratum contained 13169 trips. The non-probability sample was drawn as a simple random sample only from the large-catch stratum. In this scenario, the average catch of the large-catch subgroup was 31.89 and the average catch of the population was 27.03. When the selection mechanism depends on the response variable and can not be explained by the auxiliary information X, the strong ignorability assumption is violated. In addition, the generated non-probability sample has no chance of including a portion of the population units, resulting in under-coverage.

The estimator of $\gamma_i = P(D_i = 1|S, \mathbf{X}_i)$ was based on fitting the following model:

$$log\left(\frac{\gamma_i}{1-\gamma_i}\right) = b_0 + b_1 \times NbPassengers + b_2 \times Release + b_3 \times Hours + b_4 \times nbAnglers + \\ b_5 \times NbCrew + b_6 \times DepthPrimary + b_7 \times State + b_8 \times Weekend. \tag{4.18}$$

We acknowledge that this model may not always be the most predictive one for each replicate of the simulation, but we decided to use the same model for all scenarios for computational efficiency. In a real application, there are several steps to find the best model, such as conducting a variable selection procedure and checking the balance on covariates.

Both simulations examined 64 settings (4 probability sample sizes × 4 non-probability sample sizes × 4 non-probability sample inclusion mechanisms), with $K = 3,000$ pairs of probability sample and non-probability sample generated independently under each setting. Within each replicate, the proposed estimators $\hat{t}_{y,\ joint}$, $\hat{t}_{y,\ disjoint}$, $\hat{t}_{y,joint-adp}$, $\hat{t}_{y,disjoint-adp}$ and $\hat{t}_{y2}$ were calculated. In our population, the actual total catch $t_y = 426.38 \times 10^3$.

To aggregate the findings across all iterations, the empirical mean, variance and mean squared error (MSE) were calculated for each estimator over the $3,000$ replicates. To assess the performance of each estimator in terms of bias reduction, the relative bias of each estimator was computed as the difference between the empirical mean and the actual total catch divided by the actual total catch.

Based on the the jackknife and bootstrap variance estimates, the proportion of replicates for which the 95% confidence interval includes the actual total catch was recorded for each estimator. These proportions are denoted as $Coverage$ in the following results. The relative bias for the jackknife and bootstrap variance estimated for each estimator were also calculated as the difference between the mean of the estimated variance and the empirical variance divided by the empirical variance. A negative bias means the corresponding estimator is biased downward. These proportions are reported as percentages and denoted as $RelBias$ in the following results.

For the APSA method, the Wilcoxon rank sum test with significance level of 0.05 was used to compare the means between the two sample sources. For every replicate, the number of subgroups that have a non-significant p-value was recorded and is denoted by $\#subgroup$ in the following discussion. This is the number of subgroups that the APSA method determines to be integrated well by the PSA method. After applying the APSA method, the value of $\#subgroup$ was also calculated for the retained sample. In the simulation studies, $\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$ were computed only if the value of $\#subgroup$ from the retained sample was higher than that of the full sample.

4.6.2. Simulation Results

We present the results when the probability sample was a SRS with the self-reporting rate of 0.3. Simulation results for the complex design for this simulation settings can be found in Appendix C. For both simulation studies, the patterns of the result revealed by this reporting rate remains true for all the other self-reporting rates. In general, the performance of the five estimators are slightly worse in the complex design compared to the SRS design, which is due to the smaller effective sample sizes.

Figure 4.2 displays the empirical MSE for each estimator from all four self-reporting mechanisms: MCAR, MAR, Large Catch Case and NMAR. Table 4.2 lists the means of $\#subgroup$ for both the full sample and retained sample from the APSA method under each setting.

Table 4.2: The means of $\#subgroup$ from the APSA method when the probability sample is a SRS based on 3,000 replicates for scenarios: MCAR, MAR, Large Catch Case and NMAR.

	$\#subgroup$	Probability Sample Size (n_2)			
		200	400	600	800
MCAR	Full Sample	9.56	9.56	9.53	9.56
	Retained Sample	9.79	9.78	9.76	9.77
MAR	Full Sample	9.45	9.41	9.41	9.51
	Retained Sample	9.68	9.64	9.64	9.73
Large Catch Case	Full Sample	6.47	4.8	4.31	4.06
	Retained Sample	8.23	8.45	8.25	8.07
NMAR	Full Sample	7.07	6	5.48	5.08
	Retained Sample	7.86	7.71	7.55	7.21

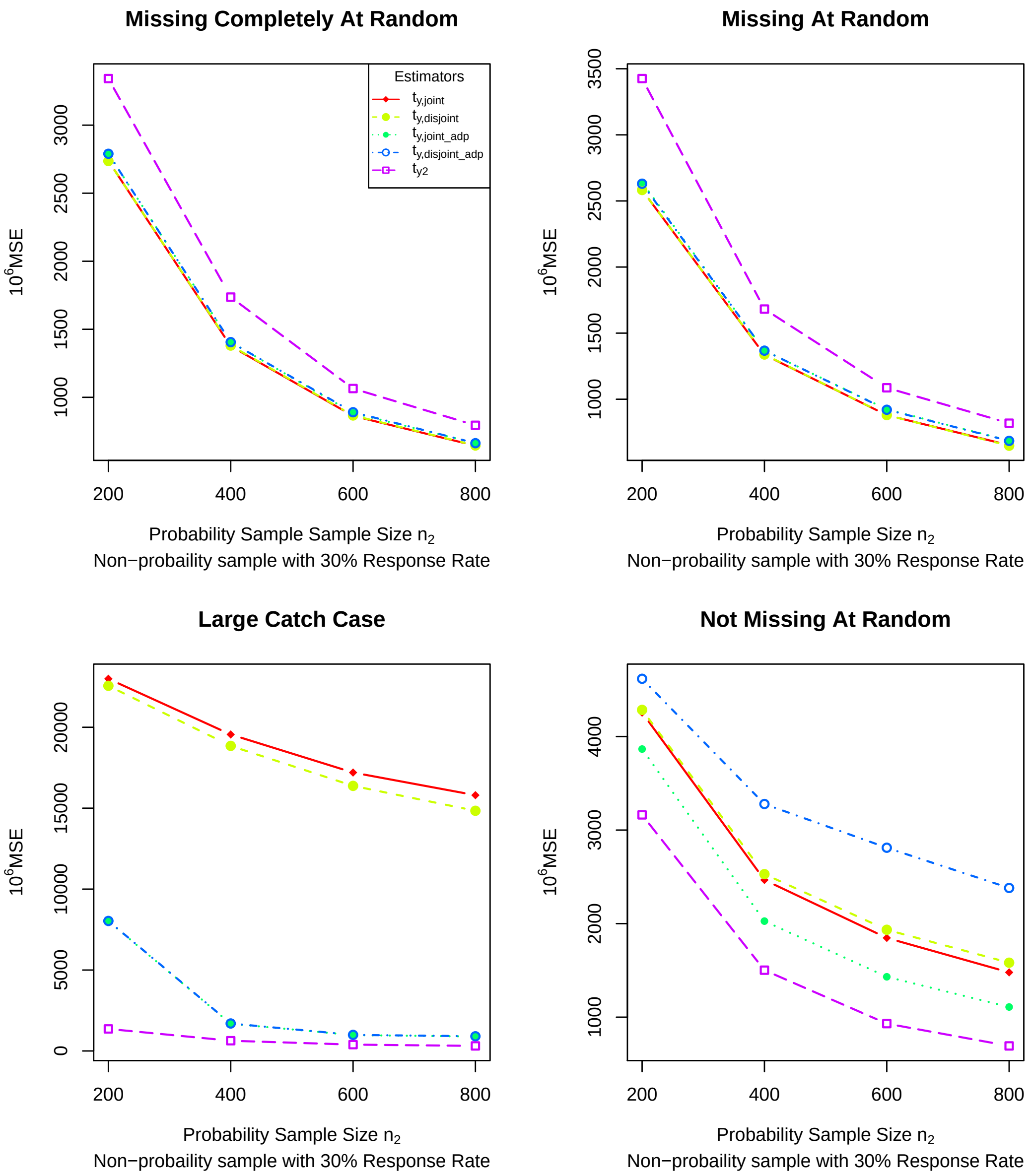


Figure 4.2: Empirical MSE of the five estimators when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for scenarios: MCAR, MAR, Large Catch Case and NMAR.

In the MCAR and MAR scenarios, $\hat{t}_{y,\ joint},\ \hat{t}_{y,\ disjoint}, \hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$ have smaller MSE compared to $\hat{t}_{y2}$. The performance of the PSA and APSA estimators are similar, which demonstrate their potential of providing improvement over $\hat{t}_{y2}$. From Table 4.2, the means of $\#subgroup$ for the full sample are above 9 in the two scenarios. This indicates that for most of the subgroups, the means of the response from the two samples within each subgroup are similar. Thus, we conclude that the two samples can be integrated well by the PSA and APSA methods. In the Large Catch and NMAR scenarios, however, both the PSA and APSA estimators have larger MSEs compared to those of $\hat{t}_{y2}$. The means of $\#subgroup$ for the full sample are around 5 in most of the settings, which indicates a model failure as the discrepancy of the response between the two samples still exist in many subgroups. Thus, there is a limit to the ability of the propensity-based pseudo-weights method to handle large selection bias. However, the means of $\#subgroup$ of the retained sample are higher than those of the full sample for all settings of each scenario, which shows that the representativeness of the retained sample can be improved by the APSA method. Similar results for the complex design are shown in Figure C.1 and Table C.1 with the same conclusion. Next, we investigate the four scenarios in detail.

Table 4.3 lists the relative bias of each estimator as a percentage for the four scenarios. In the MCAR and MAR scenarios, all the estimators are nearly unbiased. However, the non-probability sample in the MAR scenario contains moderate selection bias as the correlation between the selection probability of the non-probability sample and the response is 0.42. Thus, this scenario demonstrates the bias reduction ability of the PSA and APSA estimators when the self-reporting mechanism is correctly modeled.

In the Large Catch Case, Table 4.3 shows that the PSA estimators fail to adjust the selection bias. However, the APSA estimators remove most of the selection bias, which demonstrates their bias reduction ability compared to the PSA estimators. In the NMAR case, all PSA and APSA estimators fail to adjust the selection bias in all settings, which is due to the violation of the strong ignorability assumption and the under-coverage of the non-probability sample. As a result, the remaining selection bias becomes a major source of the inflated MSE in Figure 4.2. Compared to other estimators, $\hat{t}_{y,disjoint-adp}$ has the largest MSE with a positive relative bias for all probability sample size settings. This is because the APSA method tends to drop the non-probability sample units with low response value in this scenario, which increases the selection bias in the non-discarded part of the self-reported sample and thus be less representative of the population. However, compared to $\hat{t}_{y,disjoint-adp}$, $\hat{t}_{y,joint-adp}$ mitigates such un-representativeness and has the smallest bias and MSE for all the PSA and APSA estimators.

In addition, as showed in Table 4.2 and Table 4.3, there is a clear trend in the Large Catch Case and NMAR scenarios that when the probability sample size increases, both the relative bias and the mean of $\#subgroup$ of $\hat{t}_{y,joint-adp}$ decrease. This pattern indicates that the increased probability sample size improves the bias reduction ability of $\hat{t}_{y,joint-adp}$, which is because both the propensity score model and the Wilcoxon rank sum test become more reliable as the probability sample size increases. Similar results for the complex design are shown in Table C.2 with the same conclusion.

Table 4.4 lists the coverage rates and the relative bias of the jackknife and bootstrap variance estimates of the five estimators, along with their empirical variance for the MCAR and MAR scenarios. Overall, the bootstrap method outperforms the jackknife method and its coverage rates are close to the nominal level, 0.95, for all estimators across all settings. The coverage rates from the APSA estimators are lower than those of the PSA estimators and $\hat{t}_{y2}$, for both the jackknife and bootstrap methods, which indicates that the variance of the APSA estimators are underestimated in the two scenarios. This is because the dropping mechanism might have erroneously dropped part of the non-probability sample to make the two samples more similar, which made the estimates less variable. We also notice that the empirical variance for the PSA and APSA estimators are smaller than that of $\hat{t}_{y2}$, which demonstrates the advantage of the PSA and APSA estimators by involving a larger sample size into estimation.

Table 4.3: Relative bias of the five estimators when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for scenarios: MCAR, MAR, Large Catch Case and NMAR.

		PSA		APSA		Ratio Estimator
	n_2	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{y2}$
MCAR	200	0.01	0.01	0.01	0.01	0.01
	400	0.00	0.00	0.00	0.00	0.00
	600	0.00	0.00	0.00	0.00	0.01
	800	0.00	0.00	0.00	0.00	0.00
MAR	200	0.00	0.00	0.00	0.00	0.01
	400	0.00	0.00	0.0	0.00	0.01
	600	0.00	0.00	0.00	0.00	0.00
	800	0.00	0.00	-0.01	0.00	0.00
Large Catch Case	200	0.32	0.32	0.15	0.17	0.01
	400	0.31	0.31	0.06	0.07	0.01
	600	0.30	0.29	0.03	0.05	0.00
	800	0.29	0.28	-0.02	0.04	0.00
NMAR	200	0.09	0.09	0.08	0.12	0.01
	400	0.08	0.08	0.06	0.09	0.00
	600	0.07	0.08	0.05	0.07	0.00
	800	0.07	0.07	0.04	0.05	0.00

Table 4.5 lists the coverage rates and the relative bias of the jackknife and bootstrap variance estimates of the five estimators, along with their empirical variance for the Large Catch Case and NMAR scenarios. Due to the large selection bias, the confidence intervals based on the jackknife and bootstrap methods fail to capture the actual total catch for the PSA and APSA estimators. However, the coverage rates for the ratio estimator $\hat{t}_{y2}$ remain close to their nominal level for all the settings as expected. Similar results for the complex design are shown in Table C.6 and Table C.7 with the same conclusion.

Table 4.4: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for scenarios: MCAR and MAR.

			PSA						APSA						Ratio Estimator		
Scenario	n_2		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
MCAR	200	JK	93.52	27.28	-0.07	93.56	27.26	-0.07	92.73	27.78	-0.16	92.77	27.78	-0.16	92.28	33.3	-0.07
		Boot	94.04		0.04	94		0.03	93.84		0	93.71		0	93.16		0.04
	400	JK	93.36	13.74	-0.1	93.2	13.78	-0.11	91.87	14.02	-0.18	91.91	14.03	-0.18	92.32	17.34	-0.12
		Boot	94.32		0	94.4		-0.01	93.94		-0.03	93.81		-0.04	93.28		-0.04
	600	JK	93.68	8.61	-0.05	93.72	8.64	-0.05	92.47	8.82	-0.14	92.59	8.87	-0.14	92.96	10.59	-0.05
		Boot	94.84		0.06	94.68		0.06	94.95		0.02	94.91		0.01	94.2		0.04
	800	JK	93.72	6.44	-0.06	93.76	6.45	-0.06	92.15	6.62	-0.14	92.15	6.62	-0.14	92.72	7.93	-0.07
		Boot	94.64		0.07	94.52		0.07	94.05		0.02	94.09		0.02	94.4		0.03
MAR	200	JK	93.52	25.83	-0.05	93.68	25.83	-0.05	92.36	26.4	-0.12	92.4	26.3	-0.11	93.12	34.18	-0.11
		Boot	94.68		0.03	94.6		0.03	94.14		0.02	94.14		0.02	94		-0.01
	400	JK	93.32	13.36	-0.09	93.12	13.38	-0.09	91.51	13.68	-0.17	91.31	13.68	-0.17	93.52	16.68	-0.1
		Boot	94.28		0.01	94.24		0.01	93.13		-0.03	93.17		-0.04	93.84		0
	600	JK	92.36	8.79	-0.09	92.44	8.79	-0.09	90.39	9.23	-0.18	90.26	9.2	-0.18	93.32	10.83	-0.08
		Boot	94.04		0.01	94.28		0.01	93.16		-0.06	93.45		-0.05	94.68		0
	800	JK	93.2	6.48	-0.09	93.08	6.45	-0.08	91.38	6.82	-0.17	90.87	6.8	-0.17	92.76	8.18	-0.11
		Boot	94.64		0.03	94.64		0.03	93.72		-0.03	93.8		-0.03	94.04		-0.02

Table 4.5: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is a SRS and the reporting rate was 30% based on 3,000 replicates for scenarios: Large Catch Case and NMAR.

			PSA						APSA						Ratio Estimator		
Scenario	n_2		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
Large Catch Case	200	JK	35.2	40.35	-0.08	35.16	39.67	-0.08	66.51	40.13	-0.54	58.18	39.47	-0.56	93.92	13.54	-0.08
		Boot	41.48		0.04	42.28		0.03	76.26		-0.38	70.18		-0.37	94.6		0.03
	400	JK	4.88	18.2	-0.03	5.16	17.57	-0.03	72.09	16.03	-0.66	63.72	15.71	-0.74	94.56	6.27	-0.04
		Boot	6.96		0.09	7.36		0.09	82.66		-0.43	74.97		-0.6	95.16		0.06
	600	JK	0.4	11.09	0	0.56	10.67	-0.01	65	9.32	-0.73	54.44	8.62	-0.78	94.68	3.9	-0.01
		Boot	1.08		0.14	1.2		0.13	78.64		-0.49	71.72		-0.64	96		0.1
	800	JK	0.08	8.45	-0.06	0.12	7.97	-0.05	54.42	6.68	-0.78	46.88	6.24	-0.78	94	3.11	-0.08
		Boot	0.2		0.1	0.28		0.11	69.99		-0.53	51.64		-0.63	95.52		0.03
NMAR	200	JK	86.4	28.7	-0.06	86.32	28.6	-0.06	80.35	28.89	-0.55	76.79	29.55	-0.54	93.2	31.26	-0.09
		Boot	87.4		0.06	87.24		0.06	89.94		-0.25	87.13		-0.24	94.2		0.03
	400	JK	84.28	13.7	-0.07	83.84	13.66	-0.07	68.61	14.05	-0.66	55.17	15.17	-0.65	92.28	14.99	-0.11
		Boot	85.64		0.05	86.48		0.05	81.33		-0.39	70.31		-0.41	93.48		-0.02
	600	JK	80.48	8.75	-0.05	79.24	8.71	-0.05	61.72	9.09	-0.69	39.96	10.2	-0.69	93.68	9.27	-0.04
		Boot	83.96		0.09	82.92		0.09	77.67		-0.41	56.73		-0.44	94.92		0.06
	800	JK	76.72	6.66	-0.09	85.44	6.62	-0.09	57.15	6.95	-0.73	33.98	7.86	-0.72	92.92	6.92	-0.06
		Boot	80.28		0.05	78.32		0.06	74.48		-0.45	50.79		-0.47	94.08		0.05

From the four non-probability sample selection mechanisms, we have demonstrated that both the PSA and APSA estimators can handle moderate selection bias and the value of $^{\#}subgroup$ can serve as an indicator of whether the samples are integrated well. Obviously, a user will not know whether the data selection pattern is one of the safe settings or not. Therefore, it is still unclear how the user would know when it is safe to use estimators from APSA or PSA, and when it is not. To investigate this uncertainty, we analyzed the simulation results from a different perspective: the simulation results from the four scenarios were combined, and then classified based on $^{\#}subgroup$ from each replicate. Thus, the replicates with the same number of retained subgroups were aggregated and analyzed in the same way as previously described. The results were categorized into four groups: 10, 9, 8 and less than 8 retained subgroups. We still choose to present the results for all possible probability sample sizes when the reporting rate was 0.3. For both simulation studies, the patterns of the result revealed by this reporting rate remain true for all the other reporting rates.

Figure 4.3 displays the empirical MSE for each estimator when the numbers of retained subgroups are 10, 9, 8 and less than 8. When $^{\#}subgroup$ is 10, the MSEs of the PSA and APSA estimators are lower than those of $\hat{t}_{y2}$ for all probability sample sizes. When $^{\#}subgroup$ is 9, the APSA estimators have the smallest MSEs compared to those of the PSA estimators and $\hat{t}_{y2}$. When $^{\#}subgroup$ is less than 9, the MSEs of the PSA and APSA estimators become larger than those of $\hat{t}_{y2}$, especially when the probability sample size is small ($n_2 = 200$).

Table 4.6 lists the relative bias of each estimator as a percentage based on $^{\#}subgroup$. It indicates when the PSA and APSA estimators can handle the selection bias ($^{\#}subgroups \geq 9$) and when they can't ($^{\#}subgroups < 9$). As we discussed before, a similar trend appears for $\hat{t}_{y,joint\_adp}$ among all numbers of retained subgroups that when the probability sample size increases, its relative bias decreases. We also notice that the relative bias of the APSA estimators are smaller than those of the PSA estimators for all settings, which confirms the bias reduction ability of the APSA estimators compared to the PSA estimators. However, $\hat{t}_{y2}$ is approximately unbiased for all settings. Similar results for the complex design are shown in Figure C.2 and Table C.5 with the same conclusion.

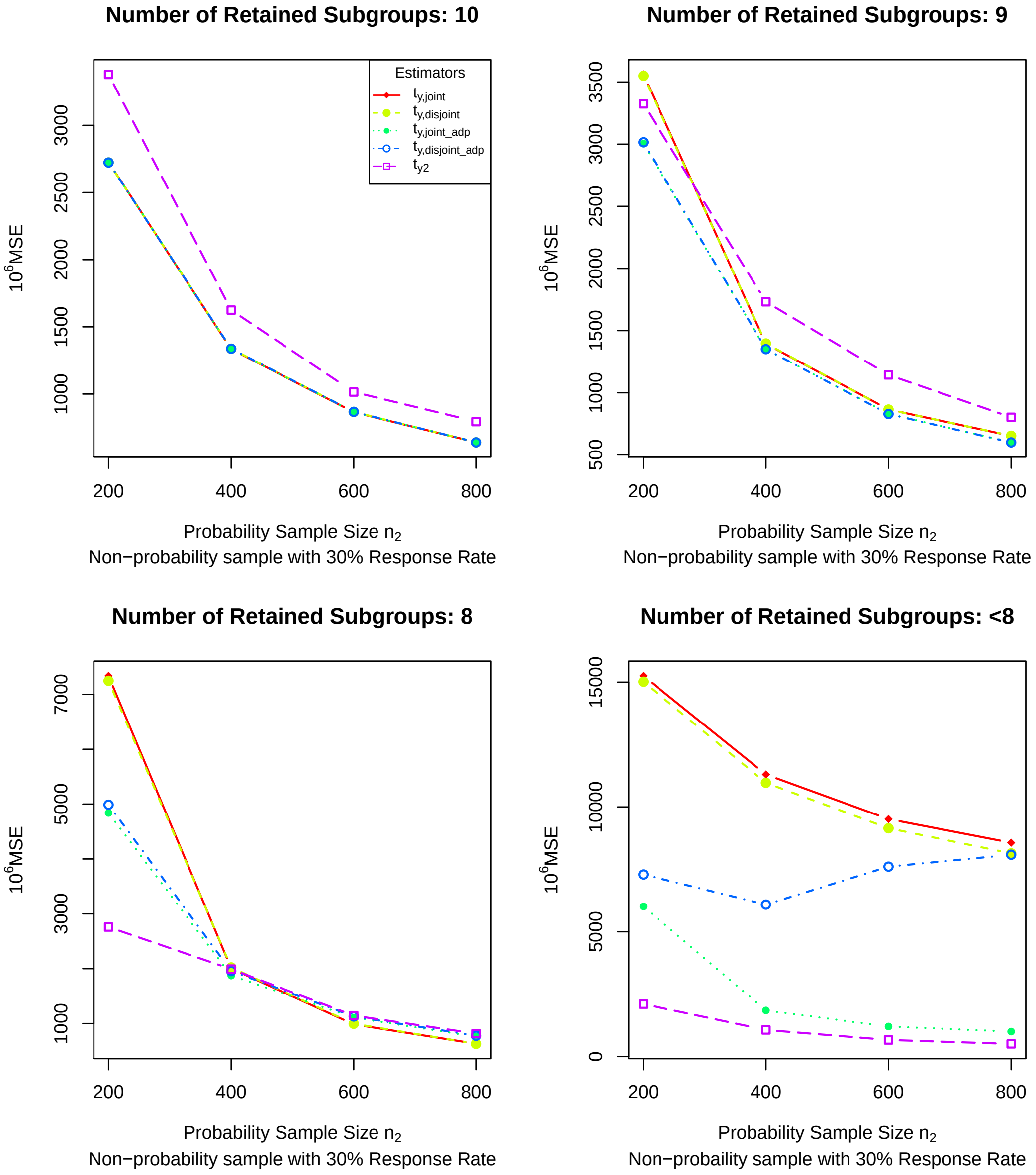


Figure 4.3: Empirical MSE of the five estimators when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for different Number of retained subgroups: 10, 9, 8 and Less Than 8.

Table 4.6: Relative bias of the five estimators when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for different number of retained subgroups: 10, 9, 8 and Less Than 8.

		PSA		APSA		Ratio Estimator
$\#subgroup$	n_2	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{y2}$
10	200	0.72	0.72	0.72	0.72	0.89
	400	0.43	0.45	0.43	0.45	0.68
	600	0.09	0.1	0.09	0.1	0.42
	800	0.04	0.04	0.04	0.04	0.26
9	200	2.38	2.38	1.26	1.23	1.03
	400	0.15	0.16	0.16	0.11	0.29
	600	0.28	0.3	0.11	0.08	0.75
	800	-0.18	-0.16	0.00	-0.02	0.18
8	200	12	11.98	8.46	8.94	1.7
	400	3.38	3.47	2.3	2.62	2.09
	600	1.49	1.52	1.54	1.62	1.31
	800	-0.16	-0.18	-0.47	-0.55	0.19
Less Than 8	200	22.19	22.07	11	13.75	0.58
	400	19.88	19.69	3.86	13.94	0.41
	600	18.47	18.27	1.79	16.35	0.37
	800	17.5	17.27	0.58	17.09	0.2

Table 4.7 and Table 4.8 list the coverage rates and the relative bias of the jackknife and bootstrap variance estimates of the five estimators, along with their empirical variance based on the number of retained subgroups. Overall, the bootstrap method outperforms the jackknife method. When $\#subgroup$ is 10, the coverage rates from the bootstrap method for all estimators are close to their nominal level, 0.95. When $\#subgroup$ is 9, the coverage rates from the bootstrap method for the APSA estimators and $\hat{t}_{y2}$ are comparable and close to 0.95. However, when $\#subgroup$ is less than 9, both the jackknife and bootstrap variance estimates fail to capture the actual variance of the PSA and APSA estimators in most of the settings, while their performance on $\hat{t}_{y2}$ are consistently reliable. From Table 4.7, it is also notable that the empirical variance of the APSA estimators are

the smallest among all settings when $\#subgroup$ is 10 or 9. As a conclusion, we suggest to use $\hat{t}_{y,joint-adp}$ instead of $\hat{t}_{y2}$ when $\#subgroup$ is greater than 8, and its variance should be estimated by the bootstrap method. Similar results for the complex design are shown in Table C.6 and Table C.7 with the same conclusion.

Table 4.7: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for number of retained subgroups: 10 and 9.

			PSA						APSA						Ratio Estimator		
$\#subgroup$	n_2		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
10	200	JK	93.72	27.16	-0.07	93.81	27.14	-0.07	93.72	27.16	-0.07	93.81	27.14	-0.07	92.94	33.65	-0.08
		Boot	94.69		0.04	94.6		0.04	94.69		0.04	94.6		0.04	93.56		0.02
	400	JK	93.85	13.3	-0.07	93.72	13.33	-0.07	93.85	13.3	-0.07	93.72	13.33	-0.07	93.66	16.16	-0.06
		Boot	94.51		0.03	94.51		0.03	94.51		0.03	94.51		0.03	94.42		0.04
	600	JK	92.71	8.65	-0.07	92.67	8.67	-0.07	92.71	8.65	-0.07	92.67	8.67	-0.07	93.88	10.12	-0.02
		Boot	94.77		0.04	94.83		0.04	94.77		0.04	94.83		0.04	94.93		0.07
	800	JK	93.67	6.4	-0.06	93.54	6.4	-0.06	93.67	6.4	-0.06	93.54	6.4	-0.06	92.8	7.93	-0.08
		Boot	94.66		0.06	94.72		0.06	94.66		0.06	94.72		0.06	94.34		0.02
9	200	JK	90.37	34.58	-0.26	89.96	34.48	-0.26	91.86	29.76	-0.3	91.75	29.88	-0.3	92.46	33.06	-0.1
		Boot	92.66		-0.18	92.45		-0.18	92.72		-0.11	92.89		-0.13	93.55		0.01
	400	JK	90.21	13.93	-0.13	90.21	13.96	-0.13	93.08	13.5	-0.28	92.88	13.5	-0.28	92.27	17.3	-0.13
		Boot	93.17		-0.03	93.01		-0.03	94.37		-0.09	94.51		-0.09	93.02		-0.06
	600	JK	91.87	8.65	-0.05	91.73	8.64	-0.05	94.28	8.28	-0.23	94.34	8.31	-0.22	93.06	11.33	-0.1
		Boot	94.6		0.05	94.6		0.05	95.05		-0.04	95.31		-0.04	93.77		-0.02
	800	JK	90.28	6.52	-0.08	90.13	6.53	-0.08	92.77	6.01	-0.24	92.84	6.14	-0.24	92.9	8.02	-0.08
		Boot	92.84		0.04	92.84		0.04	94.51		-0.05	94.18		-0.05	94.04		0.01

Table 4.8: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is a SRS and the reporting rate is 30% based on 3,000 replicates for number of retained subgroups: 8 and Less Than 8.

			PSA						APSA						Ratio Estimator		
$\#_{subgroup}$	n_2		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
8	200	JK	81.09	47.21	-0.4	81.02	46.42	-0.4	83.73	35.4	-0.46	82.45	35.38	-0.47	93.72	27.07	-0.1
		Boot	83.93		-0.33	84.16		-0.33	89.95		-0.22	89.95		-0.22	94.39		0
	400	JK	90.62	17.97	-0.32	90.62	18.05	-0.33	81.2	17.78	-0.5	79.7	18.33	-0.51	92.65	19.13	-0.24
		Boot	91.54		-0.23	92.28		-0.23	89.74		-0.31	87.39		-0.33	92.65		-0.15
	600	JK	94.04	9.44	-0.12	93.78	9.54	-0.13	85.4	10.65	-0.43	84.85	10.81	-0.44	93.52	11.13	-0.11
		Boot	95.6		-0.02	95.34		-0.03	90.36		-0.23	90.91		-0.24	95.34		-0.01
	800	JK	94.33	6.3	-0.06	94.93	6.29	-0.06	84.29	7.56	-0.41	83.01	7.72	-0.42	92.84	8.15	-0.1
		Boot	95.52		0.07	94.93		0.07	91.67		-0.19	90.71		-0.22	94.33		-0.01
Less Than 8	200	JK	60.54	63.04	-0.48	60.59	61.64	-0.47	71.03	38.16	-0.62	63.97	38.59	-0.62	93.34	20.93	-0.11
		Boot	64.99		-0.41	65.26		-0.41	81.32		-0.42	75.77		-0.42	94.29		0
	400	JK	45.71	41.24	-0.63	45.61	39.23	-0.62	70.07	15.76	-0.68	43.44	25.6	-0.75	93.23	10.61	-0.13
		Boot	47.87		-0.59	47.95		-0.57	81.65		-0.44	56.6		-0.6	94.24		-0.04
	600	JK	42.93	33.18	-0.71	42.41	30.84	-0.69	63.38	11.44	-0.77	27.2	27.5	-0.83	94.14	6.61	-0.07
		Boot	45.01		-0.67	44.55		-0.65	78.14		-0.56	39.14		-0.71	95.42		0.03
	800	JK	41.37	30.03	-0.77	40.73	27.22	-0.75	55.88	9.94	-0.83	20.83	27.79	-0.87	93.44	5.06	-0.11
		Boot	43.17		-0.73	42.24		-0.71	72.33		-0.65	31.64		-0.77	94.81		0

4.7. Case Studies

The purpose of this example is to illustrate when the proposed PSA and APSA estimators can be successful alternatives to the ratio estimator $\hat{t}_{y2}$. The dockside intercept sample (MRIP) and the self reported sample (ELB) used in this example came from an ELB experiment conducted in Florida, which was described in Section 3.3. The dockside intercept sample (MRIP) contained 1142 charter trips and the self reported sample (ELB) contained 4649 charter trips. Among the two samples, there were 86 matched trips. We assume the self reports were accurate and every boat was able to self report. Since neither sample had clear PSU information available to us, we defined the PSU as described in Subsection 4.6.1.

The PSA and APSA methods were applied to integrate the two samples and estimate the total catch of two fish species: Red Snapper and Red Porgy. They represent two types of fishing targets whose total abundance are of interest to NOAA. Red Snapper is a high value and commonly targeted species in Gulf of Mexico. Since the late 1990s, annual recreational landings of this fishery have passed 5 million pounds, while it was less than 500,000 pounds prior to 1950. Certain regulations on fishing season and bag limits have been implemented to protect this population from overfishing. By contrast, Red Porgy is often caught by recreational anglers in the Gulf of Mexico as an off-target species. Few regulations has been set on this population.

4.7.1. Summary Statistics

The differences in nature, value and regulations between the two fish species result in different catch distributions. Table 4.9 lists the summary statistics of catch for the two species. For both species, the percentages of trips with landings, average catch and variance from the dockside intercept sample are much lower than those from the self-reported sample. By contrast, the CVs from the dockside intercept sample are larger than those from the self-reported sample. This shows that anglers tend to report their trip when they have a harvest. Compared to Red Snapper, the difference of catch distributions between the two samples for Red Porgy is smaller. This indicates the self reports on Red Porgy is more representative of its fishery population.

Table 4.9: Summary statistics of the MRIP sample and ELB sample for the Red Snapper and Red Porgy catches.

	Red Snapper		Red Porgy	
	MRIP	ELB	MRIP	ELB
Percentage of the trips with landings	0.07	0.38	0.11	0.27
Mean of catch	0.58	4.61	1.29	3.41
Variance of catch	7.72	63.55	32.57	70.75
CV	4.79	1.72	4.42	2.46

The auxiliary information from the two samples were also explored to investigate the difference. Table 4.10 shows the 5 variables that were used to build the propensity score model. The numbers of released Red Snapper and Red Porgy were used separately to build their own propensity score models. Statistical tests (wilcoxon wanked sum test for non-categorical variables and χ^2 test for categorical variables) were conducted for every variable to decide whether differences between the two samples are significant.

Table 4.10 shows a considerable difference in all variables between the two samples. For example, the average number of anglers from the self-reported sample is higher than the dockside intercept sample, which may indicate that larger boats are more likely to report their trip information. Among the self-reported sample, 74% of the trips targeted valuable fish species, while this proportion is only 16% for the dockside intercept sample. This suggests that anglers tend to report more on purposive trips rather than recreational only. It is also notable that the numbers of released Red Porgy were small from both samples. This could be due to either of the following two reasons: 1) anglers tend to keep Red Porgy if they catch one since they are not regulated, 2) anglers tend not to make note of how many Red Porgy they released, so nothing is recorded, and thus 0 is imputed.

Table 4.10: Covariates of the propensity score model for the total Red Snapper and Red Porgy catches by Charter boat.

Variable Name	Type	Description	Mean		P-value
			MRIP	ELB	
Number of Anglers	Cont.	Number of anglers on the boat	3.87	6.05	0.00
Weekend	Cat.	1: The trip is made on weekend 0: The trip is made on weekday	0.42	0.29	0.00
Hours	Cont.	Fishing duration	4.03	3.29	0.00
Target	Cat.	1: The trip targets on Red Snapper, Vermilion Snapper, King Mackerel and Red Grouper. 0: Otherwise	0.16	0.74	0.00
Red Snapper Released	Cont.	Alive Red Snapper fish released	1.52	7.39	0.00
Red Porgy Released	Cont.	Alive Red Porgy fish released	0.00	0.43	0.00

4.7.2. The PSA Method

To assess how the decision to report could be explained by these variables, the propensity score models were built on the combined sample to predict the probability for a trip to be reported. The models for Red Snapper and Red Porgy shared the same covariates; $NumberofAnglers$, $Weekend$, $Hours$ and $Target$, while the number of released Red Snapper was included in the model for Red Snapper and the number of released Red Porgy was included in the model for Red Porgy. The resulting propensity scores were used to calculate sample weights for the PSA method.

The model estimates for Red Snapper and Red Porgy are tagged as PSA in Table 4.11. All variables are significant at the 5% level. The estimates are similar from both models, and they illustrate a similar reporting manner as noted in Table 4.9: A trip is more likely to be self reported with a clear fishing target and when it involves more anglers. In addition, both models indicate that a trip is more likely to be self reported when it is made on a weekday with a shorter duration, and with a harvest as reflected by the larger number of fish released.

Table 4.11: Coefficient estimates of the propensity score models from the PSA and APSA methods for the total Red Snapper and Red Porgy catches by Charter boat.

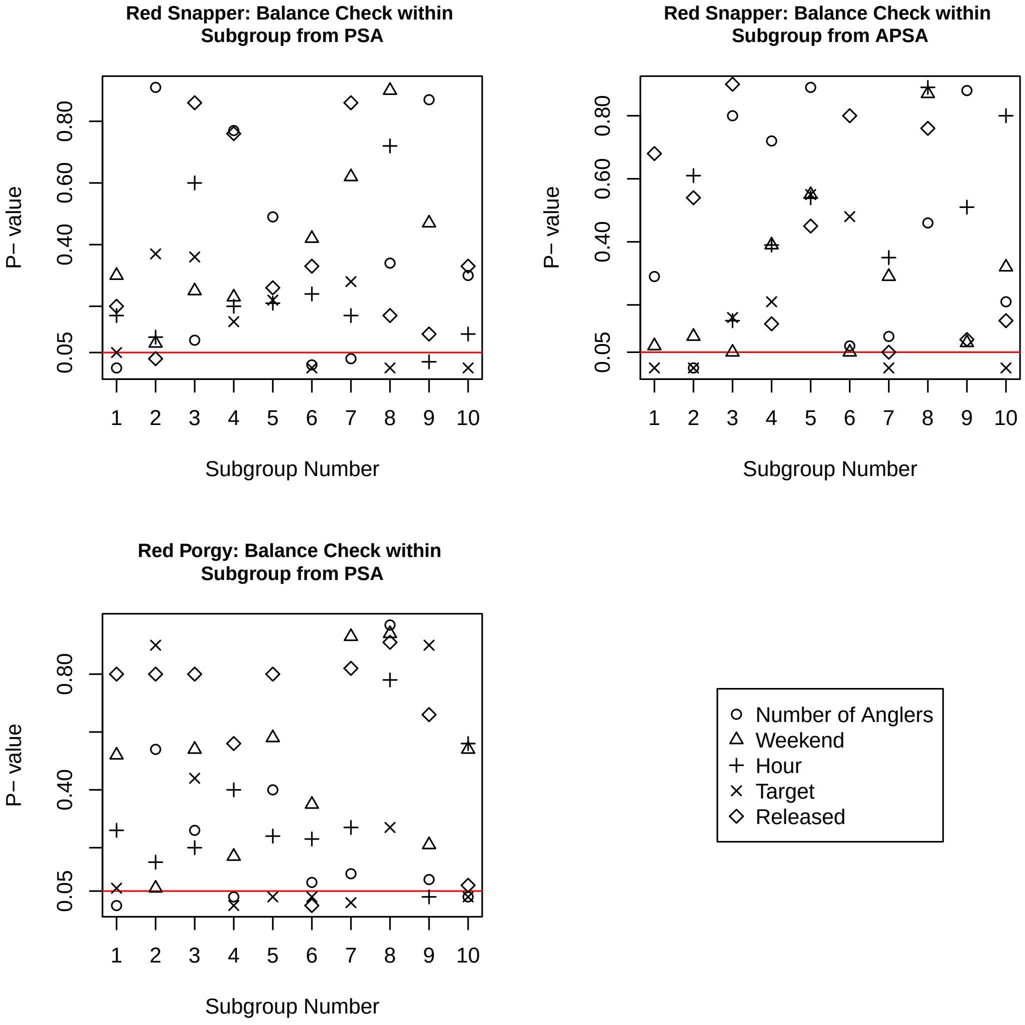
Covariate		Red Snapper		Red Porgy
		PSA	APSA	PSA
		Odds Ratio (95% CI)	Odds Ratio (95% CI)	Odds Ratio (95% CI)
Number of Anglers		1.40 (1.33, 1.46)	1.57 (1.48, 1.63)	1.43 (1.36, 1.50)
Weekend	Weekday			
	Weekend	0.63 (0.53, 0.74)	0.50 (0.39, 0.64)	0.62 (0.52, 0.74)
Hours		0.79 (0.77, 0.83)	0.78 (0.73, 0.83)	0.82 (0.79, 0.85)
Target	Off-target Species			
	Target Species	11.05 (9.11, 13.47)	65.79 (51.57, 84.67)	12.29 (10.17, 14.95)
Red Snapper Released		1.05 (1.04, 1.07)	1.09 (1.07, 1.10)	
Red Porgy Released				4.33 (1.69, 58.67)

4.7.3. Balance Assessment

To assess whether homogeneous trips from the two samples have been successfully grouped by the propensity score model, the combined sample was segmented into 10 subgroups based on the estimated propensity score and the balance of covariates were checked across all subgroups by statistical tests. The subgroups were defined as described in Subsection 4.4.2 for the APSA method. The Wilcoxon ranked sum test was used for non-categorical variables and χ^2 test was used for categorical variables. Fisher's exact test was used as the alternative to χ^2 test when it was not applicable due to small category size.

Figure 4.4 shows the p-values of every covariate across every subgroup from the propensity score models for the two species. Overall, all covariates except $Target$ are balanced across all subgroups. The variable $Target$ is not balanced in many subgroups: 6, 8 and 10 for Red Snapper and 4, 5, 6, 7 and 10 for Red Porgy. This may due to the large proportion of trips with clear fishing targets in the self-reported sample, which makes the adjustment unsuccessful.

Figure 4.4: Balance check for covariates from the PSA and APSA methods for Red Snapper and the PSA method for Red Porgy.


Red Snapper: Balance Check within Subgroup from PSA

Red Snapper: Balance Check within Subgroup from APSA

Red Porgy: Balance Check within Subgroup from PSA

4.7.4. The APSA Method

Next we applied the APSA method to estimate catch for the two species, using the estimated propensity scores as described in the previous section. For Red Snapper, Table 4.12 lists the numbers of trips from both samples, their mean catches and the p-value of the Wilcoxon rank sum test for each subgroup. Figure 4.5 shows the catch distributions within each subgroup from both samples. Table 4.12 shows that most of the MRIP trips are from the first 3 subgroups, which include trips that are less likely to be reported. By contrast, the self-reported trips dominate subgroups 5 to 10, which contain trips that are more likely to be reported. The mean catch increases from subgroups 1 to 10 among both samples, which indicates that a trip is more likely to be reported when its fish catch is larger. Similar data for Red Porgy are shown in Table D.1 and Figure D.1 with the same conclusion, which can be found in Appendix D.

Table 4.12 also shows that the catch distributions of Red Snapper are significantly different between the two samples for the first three subgroups. Figure 4.5 shows that while most of the trips in the three subgroups have zero Red Snapper catch, the self-reported sample contains relatively more trips with non-zero catch, which causes the difference. Red Porgy, by contrast, has similar catch distributions in the two samples across all subgroups, which indicates a good balance on the catch distributions between the two samples. This can be seen from Table D.1 and Figure D.1 in Appendix D.

Table 4.12: Numbers of trips for Red Snapper from the MRIP and ELB samples within each subgroup based on the PSA method.

Subgroup	Number of	Number of	Mean Catch		P-value
Number	MRIP Trips	ELB Trips	MRIP	ELB	
1	469	116	0.01	0.51	0.00
2	236	325	0.07	0.33	0.02
3	138	428	0.38	1.15	0.04
4	80	494	1.43	1.78	0.22
5	44	534	1.75	2.32	0.59
6	40	521	2.08	2.72	0.88
7	22	550	4.32	4.32	0.97
8	12	555	5.83	6.02	0.88
9	7	563	9.29	7.86	0.50
10	8	563	9.25	12.63	0.38

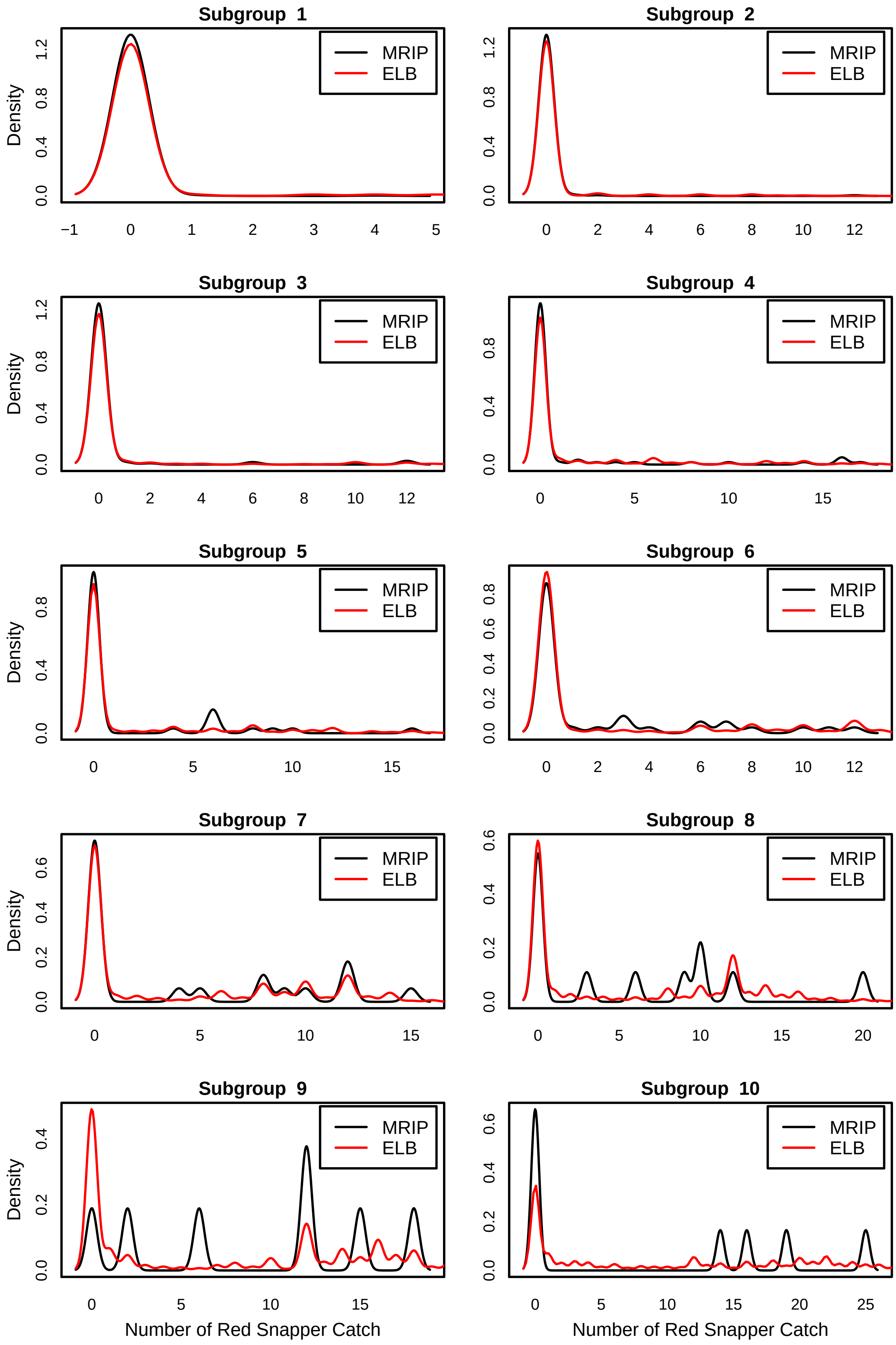


Figure 4.5: Distributions of the Red Snapper catch from the MRIP and ELB samples within each subgroup based on the PSA method.

To estimate the total catch for Red Snapper by the APSA method , we dropped the self-reported sample from the first three subgroups and recalculated the propensity score for the retained sample. Table 4.11 shows the new model estimates listed under APSA. The estimates of covariates are similar to those of the original model except the variable $Target$. The estimate of odds ratio for $Target$ increased from 11.05 to 65.79. This is because the proportion of trips with a clear target species among the dropped self-reported sample is only 1.2%, which is much lower than 74% of the full self-reported sample. Consequently, the proportion of trips with clear a target species is even higher in the retained self-reported sample. This strengthens the variable's predictability. Figure 4.4 shows the result of a balance check on covariates from the new model. Not surprisingly, almost all covariates are balanced well across the subgroups except the variable $Target$. Table 4.13 lists the number of trips from both samples, the mean catch and the P-value of the Wilcoxon rank sum test within each subgroup based on the new estimated propensity scores. After dropping the self-reported sample from the first three subgroups, the catch distributions of Red Snapper are similar between the two samples across all subgroups. Such a balance is also confirmed by Figure 4.6. For Red Porgy, since the catch distributions between the two samples were already balanced across all subgroups based on the PSA method, no self-reported trip was dropped.

4.7.5. Estimation Results

Table 4.14 shows the estimates of total from both the PSA method, the APSA method and the ratio estimator $\hat{t}_{y2}$, along with estimates of standard error from the jackknife and bootstrap estimators. The bootstrap variance estimation was based on 100 replicates. For Red Snapper, $t_{y,joint\_adp}$ provides a very close estimate compared to the ratio estimator $\hat{t}_{y2}$, while its jackknife and bootstrap standard deviations are slightly higher than theose of $\hat{t}_{y2}$. It is also notable that even though the APSA estimators employs a smaller sample size compared to the PSA estimators, their jackknife and bootstrap standard errors are

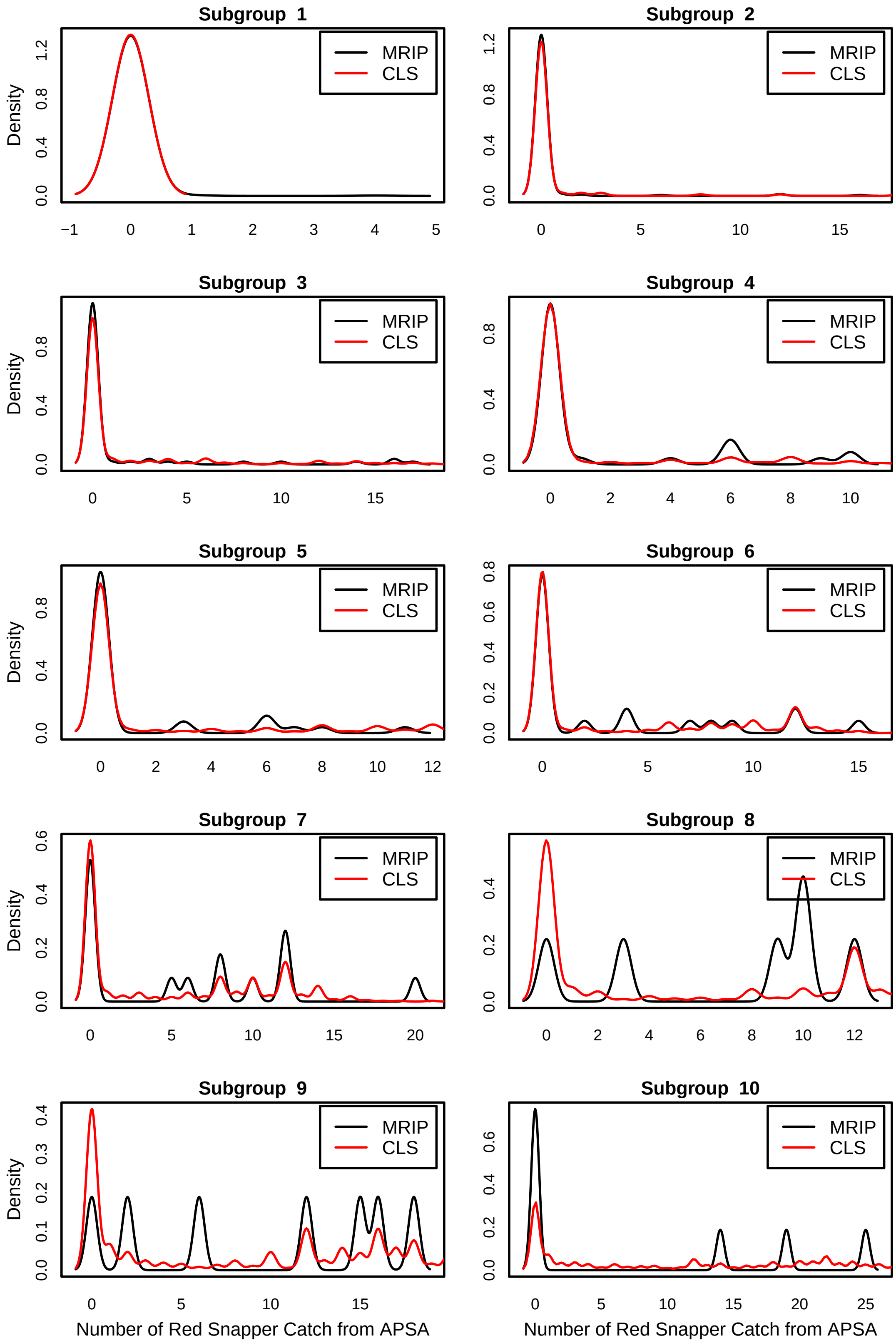


Figure 4.6: Distributions of the Red Snapper catch from the MRIP and ELB samples within each subgroup based on the APSA method.

Table 4.13: Numbers of trips for Red Snapper from the MRIP and ELB samples within each subgroup based on the APSA method.

Subgroup Number	Number of MRIP Trips	Number of ELB Trips	Catch Mean MRIP	Catch Mean ELB	P-value
1	485	7	0.01	0.00	0.84
2	373	111	0.27	0.45	0.08
3	70	424	1.41	1.86	0.24
4	35	446	1.66	2.06	0.83
5	36	498	1.39	2.33	0.40
6	22	425	3.32	3.68	0.94
7	15	469	6.20	5.38	0.68
8	6	479	7.33	6.33	0.72
9	7	481	9.86	8.86	063
10	7	481	8.29	13.50	0.21

smaller than those of the PSA estimators. This indicates that the self-reported sample is better integrated with the MRIP sample by the APSA method. However, by losing the advantage of the larger sample size, the estimators from PSA and APSA have larger variance compared to $\hat{t}_{y2}$ in this case and would not be preferred.

For Red Porgy, since no trip was dropped from the self-reported sample, $\hat{t}_{y,joint}$ and $\hat{t}_{y,joint\_adp}$ provide the same estimates, and they are very close to $\hat{t}_{y2}$. The jackknife and bootstrap standard errors for all the PSA and APSA estimators are smaller than those of $\hat{t}_{y2}$. Based on the simulation results as we discussed earlier, we recommend to use $\hat{t}_{y2}$ for Red Snapper and to use $\hat{t}_{y,joint}$ or $\hat{t}_{y,joint\_adp}$ for Red Porgy in this example. We also recommend to used the bootstrap method for variance estimation.

Table 4.14: Estimates of total by the PSA estimators, APSA estimators and t_{y2} for Red Snapper and Red Porgy by Charter boat from the MRIP Sample and ELB Sample.

	PSA		APSA		
Red Snapper	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{y2}$
Estimate ($\times 10^3$)	61.73	77.32	64.65	71.28	64.38
SE(jackknife) ($\times 10^6$)	14.67	20.57	12.59	16.44	12.30
SE(bootstrap) ($\times 10^6$)	22.29	33.45	18.60	25.17	18.62
Red Porgy	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{y2}$
Estimate ($\times 10^3$)	123.16	135.01	123.16	135.01	123.80
SE(jackknife) ($\times 10^6$)	26.15	33.85	26.15	33.85	37.93
SE(bootstrap) ($\times 10^6$)	45.37	51.99	45.37	51.99	56.72

4.8. Discussion and Future Research Plans

Based on the simulation results, both the PSA and APSA estimators have shown their potential of being useful alternatives to the current ratio estimator $\hat{t}_{y2}$. However, they have certain limits in handling the selection bias. Compared to the PSA method, the APSA method can reduce the selection bias by comparing the two samples based on the propensity score, detecting and leaving out the non-representative part of the non-probability sample. We also conclude that the joint weighting approach is more reliable than the disjoint weighting approach. In addition, the performance of the APSA method will benefit from a larger sample size of the probability sample. The number of retained subgroups from the APSA method can serve as an indicator of whether the PSA and APSA estimators are better than the current ratio estimators. As a conclusion, we recommend to use $\hat{t}_{y,joint\_adp}$ instead of $\hat{t}_{y2}$ when the number of retained subgroups is greater than 8, and the bootstrap variance estimation method is preferable.

However, there are some concerns of generalizing the PSA and APSA methods in our application. As showed in the case studies, the PSA and APSA methods are not as straightforward to use as the ratio estimators are. Even though we have specified the situations when they can be successful alternatives to the ratio estimators, it raises the time concern for NOAA to apply the new methods for every species. Since NOAA is responsible to estimate hundreds of fish species across different waves, areas and modes, it is time consuming to build the propensity score model for every single case. However, there are many situations when the ratio estimators couldn't provide precise estimates. For example, in Chapter 3 we showed that the PSEs of the ratio estimators will be increased for the species with large CV, small dockside intercept sample size and low reporting rate. Thus, we recommend NOAA to consider using the PSA and APSA estimators for such situations first.

There is also much work that needs to be done to make this research complete. First, in this research we have assumed the self reports are accurate. Unfortunately, this is not true in practice. As we observed from the matched trips from the dockside intercept sample and the self reports, there are a considerable number of trips whose catch are different from the two sample sources. There are two reasons that may cause this difference. The first reason is due to the non-sampling error that the two trips are actually mismatched. Stokes et al. (2019a) studied the effect of non-sampling error on the ratio estimators. However, it is still unclear how the non-sampling error will affect the PSA and APSA estimators. The second reason is due to the measurement error contained in the self reports. One way to address this issue is to study the relationship between the actual catch and the reported catch from the overlap of the two samples, and then predict the actual catch based on the self-reported catch for the rest of the self-reported sample. The prediction variance can be estimated by multiple imputation. However, by doing so, the variance of the PSA and APSA estimators will be greatly increased, which compromises their advantage of having a larger sample size. This is because the prediction model is built on the matched trips, whose sample size is much smaller than the self-reported sample. So other ways of dealing with the measurement error need to be studied for the PSA and APSA methods.

Second, in this study we assumed every trip in the population has a positive chance to be self reported. However, in practice the self-reported sample has the coverage issue that only the boat with the self-reporting device installed has the chance to report. The coverage issue can be ignored if the boat with the device is representative of the population. In our case studies, this assumption was met for Red Porgy even though it was not clearly stated. This is because the propensity model balanced the two sample sources well without considering the device factor. However, if this is not true, as the Red Snapper case, we can partition the population into two domains based on whether the boat has the self reporting device and make separate total estimation for each domain. In some applications, such as charter boats, this is known for the entire fleet.

Third, the sensitivity of our estimators to model mis-specification needs more scrutiny. The model mis-specification can be caused by two reasons: 1) the model doesn't include the relevant variables, 2) the model includes irrelevant variables. In the NMAR scenario of our simulation studies, we studied the effect of the first reason. However, our simulation didn't consider the effect of the second reason, as we used the same model for all scenarios. As a result, it is unclear how much bias occurs from a mis-specified model. We would expect that the performance of both the PSA and APSA estimators be improved in our simulation if the model was correctly specified in each replicate.

For the relative sample size issue, the simulation results show that the performance of the APSA methods will benefit from a larger sample size of the probability sample. However, for a given non-probability sample size, it is unclear how large a probability sample size is required for the APSA method to provide a reliable estimate. As showed in Figure 4.3, when the number of retained subgroups is greater than 8, the performance of $\hat{t}_{y,joint\_adp}$ is consistently better than the ratio estimator $\hat{t}_{y2}$ regardless of the probability sample size. However, when the number of retained subgroups is 8 and the probability sample size is small ($n_2 = 200$), $\hat{t}_{y2}$ outperforms $\hat{t}_{y,joint\_adp}$. This may not be problematic so far for the cases we studied, as the ratio of the probability sample size over the non-probability sample size was 1142/4649 = 0.24, which was close to the largest ratio in the simulation settings (800/3154 = 0.25). Under this ratio, $\hat{t}_{y,joint-adp}$ showed good performance from the simulation results. However, this problem needs to be addressed when the self-reporting rate is getting higher in the future. On the other hand, if the self-reporting rate is high enough, the non-probability sample will resemble a simple random sample and adjustment will be less important.

Fourth, it is worth investigating whether machine learning models (such as k-nearest neighbors, support vector machine) can be successful alternatives to the traditional propensity score model that used in our study. They have shown to be able to remove selection bias more efficiently than logistic regression when used for PSA (Ferri-García and Rueda, 2020). However, the machine learning models require a very large training data set to guarantee their prediction accuracy. In addition, the new models may not be able to indicate a model failure, as the APSA method does.

Last, the idea of the APSA method can be applied to many non-probability sampling applications, as it offers a way to indicate whether the non-probability sample is being correctly used. For the situations when the response variable is not available for the probability sample, one could try to apply the APSA method on the covariates that is highly correlated with the response variable.

APPENDIX A

The Mean, Variance and CV of the New Self-reported Sample

Under the assumptions discussed in Section 3.1, the relationships between p_1 and $\bar{y}_1$, CV_{1y}, $\bar{y}_1^*$, CV_{1y^*} can be specified as follows. When the reporting rate (p_1) decreases from the current level, the values of $\bar{y}_1$, CV_{1y}, $\bar{y}_1^*$ and CV_{1y^*} does not change. When the reporting rate increases by δ, denote the mean and variance of the target population, the self-reported sample, the anglers who do not report and the new self-reported sample by ($\bar{y}$, S_y^2), ($\bar{y}_1$, S_{1y}^2), ($\bar{y}_{1,c}$, $S_{1y,c}^2$) and ($\bar{y}_{1,p_1+\delta}$, $S_{1y,p_1+\delta}^2$), respectively. By the following relationships:

$$\bar{y} = p_1\bar{y}_1 + (1-p_1)\bar{y}_{1,c}, \tag{A.1}$$

and

$$S_y^2 = p_1 S_{1y}^2 + (1-p_1)S_{1y,c}^2 + p_1(1-p_1)(\bar{y}_1 - \bar{y}_{1,c})^2, \tag{A.2}$$

we could get:

$$\bar{y}_{1,c} = \frac{\bar{y} - p_1\bar{y}_1}{1-p_1} \tag{A.3}$$

and

$$S_{1y,c}^2 = \frac{S_y^2 - p_1 S_{1y}^2}{1-p_1} - p_1(\bar{y}_1 - \bar{y}_{1,c})^2. \tag{A.4}$$

Then the mean and variance for the new self-reported sample have the expressions:

$$\bar{y}_{1,p_1+\delta} = \frac{p_1\bar{y}_1 + \delta\bar{y}_{1,c}}{p_1+\delta}, \tag{A.5}$$

and

$$S^2_{1y,p_1+\delta} = \frac{p_1 S^2_{1y} + \delta S^2_{1y,c}}{p_1+\delta} + \frac{p_1\delta(\bar{y}_1 - \bar{y}_{1,c})^2}{(p_1+\delta)^2}. \quad \text{(A.6)}$$

As a result,

$$CV_{1y,p_{1+\delta}} = \frac{S_{1y,1+\delta}}{\bar{y}_{1,p_1+\delta}}. \quad \text{(A.7)}$$

Similarly, when the self reports contains measurement error, the average catch among the new self-reported sample is:

$$\bar{y}^*_{1,p_1+\delta} = \frac{p_1\bar{y}^*_1 + \delta\bar{y}^*_{1,c}}{p_1+\delta}, \quad \text{(A.8)}$$

where $\bar{y}^*_{1,c} = \frac{\bar{y}-p_1\bar{y}^*_1}{1-p_1}$ is the average catch among the anglers who do not report. Denote the variance of catch of the original self-reported sample by $S^2_{1y^*}$, then the variance of catch among the new self-reported sample is:

$$S^2_{1y^*,1+\delta} = \frac{p_1 S^2_{1y^*} + \delta S^2_{1y^*,c}}{p_1+\delta} + \frac{p_1\delta(\bar{y}^*_1 - \bar{y}^*_{1,c})^2}{(p_1+\delta)^2}, \quad \text{(A.9)}$$

where $S^2_{1y^*,c} = \frac{S^2_y - p_1 S^2_{1y^*}}{1-p_1} - p_1(\bar{y}^*_1 - \bar{y}^*_{1,c})^2$ is the variance of catch among the anglers who do not report. As a result,

$$CV_{1y^*,p_{1+\delta}} = \frac{S_{1y^*,1+\delta}}{\bar{y}^*_{1,p_1+\delta}}. \quad \text{(A.10)}$$

APPENDIX B

Large Sample Variance of $\hat{t}_{SRS}$

We derive the variance of

$$\hat{t}_{SRS} = \hat{N}\bar{y}_S = n_1 n_2 \frac{\bar{y}_S}{m} \tag{B.1}$$

discussed in Section 4.2. In the simplified case, both the non-probability sample (S_1) and the probability sample (S_2) are simple random samples with sample sizes n_1 and n_2, respectively. The random components in equation B.1 is the overlap size m and the sample average $\bar{y}_S$.

First, the overlap size m follows a hypergeometric distribution with parameters $N, n_1, and\, n_2$, so

$$E(m) = \frac{n_1 n_2}{N} \; and \; Var(m) = \frac{n_1 n_2 (N - n_1)(N - n_2)}{N^2(N-1)}. \tag{B.2}$$

Second, as $\bar{y}_S$ is the sample average from a simple random sample with size $n_S = n_1 + n_2 - m$, its variance has the expression:

$$\begin{aligned} Var(\bar{y}_S) &= E(Var(\bar{y}_S|m)) + Var(E(\bar{y}_S|m)) \\ &= E((1 - \frac{n_1 + n_2 - m}{N})\frac{S_y^2}{n_1 + n_2 - m}) + Var(\bar{y}) \\ &= S_y^2 E(\frac{1 - \frac{n_1+n_2-m}{N}}{n_1 + n_2 - m}) + 0 \\ &\approx S_y^2 (\frac{1 - \frac{n_1+n_2-E(m)}{N}}{n_1 + n_2 - E(m)}). \end{aligned} \tag{B.3}$$

Next, we have:

$$\begin{aligned} Cov(m, \bar{y}_S) &= E(Cov(m, \bar{y}_S|m)) + Cov(E(m|m), E(\bar{y}_S|m)) \\ &= 0. \end{aligned} \tag{B.4}$$

The variance of $\hat{t}_{SRS}$ can be approximated using Taylor linearizaiton when sample size is large:

$$\begin{aligned} Var(\hat{t}_{SRS}) &= Var(n_1 n_2 \frac{\bar{y}_S}{m}) \\ &= n_1^2 n_2^2 Var(\frac{\bar{y}_S}{m}) \\ &\approx n_1^2 n_2^2 \{(\frac{1}{E(m)})^2 Var(\bar{y}_S) + (-\frac{\bar{y}}{(E(m))^2})^2 Var(m) - 2\frac{\bar{y}}{(E(m))^3} Cov(m, \bar{y}_S)\} \\ &\approx N^2 \{S_y^2 (\frac{1 - \frac{n_1+n_2-E(m)}{N}}{n_1 + n_2 - E(m)}) + \frac{\bar{y}^2}{n_1 n_2} \frac{(N-n_1)(N-n_2)}{N-1}\} \\ &\approx \frac{N^2}{E(n_S)} \{S_y^2 (1 - \frac{E(n_S)}{N}) + \bar{y}^2 \frac{E(n_S)}{p_1 N n_2} (1-p_1)(N-n_2)\} \\ &\approx \frac{N^2}{E(n_S)} \{S^2 (1 - \frac{E(n_S)}{N}) + \bar{y}^2 \frac{E(n_S)}{n_2} \frac{(1-p_1)}{p_1} (1 - \frac{n_2}{N})\}, \end{aligned} \tag{B.5}$$

where $E(n_S) = n_1 + n_2 - \frac{n_1 n_2}{N}$.

APPENDIX C

Simulation Results from the Complex Design

Table C.1: The mean of $^{\#}Subgroup$ based on the APSA method when the probability sample is drawn according to a complex design based on 3,000 replicates for scenarios: MCAR, MAR, Large Catch Case and NMAR.

	$^{\#}subgroup$	Probability Sample Size (n_{PSU})			
		30	40	50	60
MCAR	Full Sample	9.36	9.36	9.36	9.37
	Retained Sample	9.59	9.59	9.6	9.6
MAR	Full Sample	9.38	9.32	9.33	9.34
	Retained Sample	9.61	9.56	9.59	9.56
Large Catch Case	Full Sample	6.19	4.97	4.55	4.3
	Retained Sample	8.13	8.22	8.09	7.94
NMAR	Full Sample	6.61	5.73	5.34	4.97
	Retained Sample	7.67	7.45	7.3	6.98

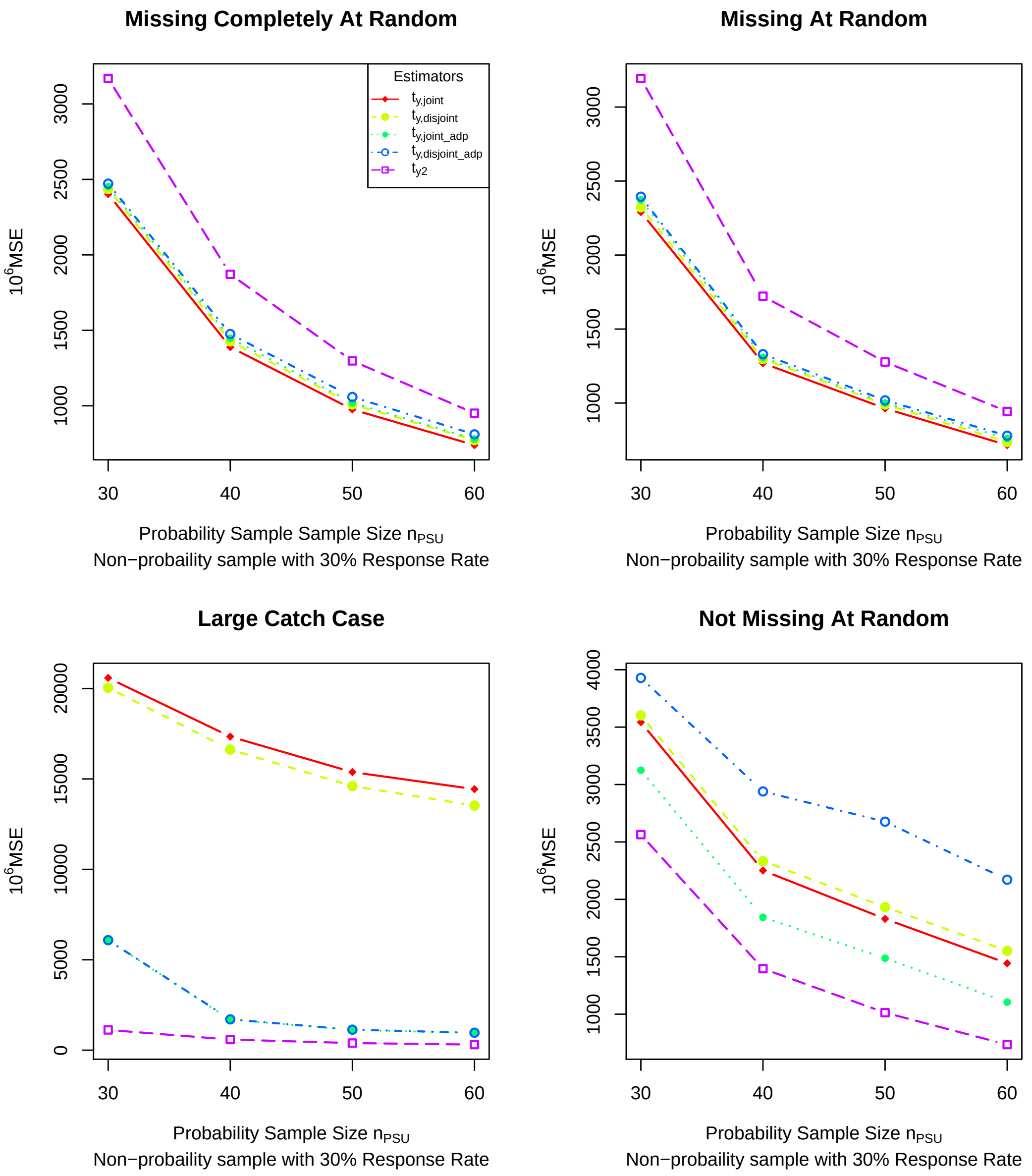


Figure C.1: Empirical MSE of the five estimators when the probability sample is according to a complex design and the reporting rate is 30% based on 3,000 replicates for scenarios: MCAR, MAR, Large Catch Case and NMAR.

Table C.2: Relative bias of the five estimators when the probability sample is according to a complex design and the reporting rate is 30% based on 3,000 replicates for scenarios: MCAR, MAR, Large Catch Case and NMAR.

		PSA		APSA		Ratio Estimator
	n_{PSU}	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{y2}$
MCAR	30	0.01	0.01	0.01	0.01	0.01
	40	0.00	0.00	0.00	0.00	0.01
	50	0.00	0.00	0.00	0.00	0.01
	60	0.00	0.00	0.00	0.00	0.00
MAR	30	0.00	0.00	0.00	0.00	0.01
	40	0.00	0.00	0.00	0.00	0.01
	50	0.00	0.00	0.00	0.00	0.00
	60	0.00	0.00	-0.01	-0.01	0.00
Large Catch Case	30	0.31	0.3	0.11	0.16	0.01
	40	0.29	0.29	0.02	0.04	0.00
	50	0.28	0.27	-0.01	0.02	0.00
	60	0.27	0.26	-0.03	0.02	0.00
NMAR	30	0.08	0.08	0.06	0.09	0.01
	40	0.08	0.08	0.05	0.09	0.00
	50	0.07	0.08	0.05	0.09	0.00
	60	0.07	0.07	0.04	0.08	0.00

Table C.3: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is drawn according to a complex design and the reporting rate is 30% based on 3,000 replicates for scenarios: MCAR and MAR.

			PSA						APSA						Ratio Estimator		
Scenario	n_{PSU}		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
MCAR	30	JK	93.4	23.94	-0.06	93.36	24.26	-0.06	92.07	24.44	-0.17	92.07	24.64	-0.16	91.76	31.54	-0.07
		Boot	92.68		-0.06	92.68		-0.05	92.37		-0.08	92.16		-0.08	91.44		-0.07
	40	JK	93.24	13.86	-0.07	93.16	14.23	-0.07	91.59	14.44	-0.18	91.63	14.71	-0.18	92.28	18.64	-0.08
		Boot	93.6		-0.06	93.6		-0.06	93.08		-0.09	93.16		-0.09	91.96		-0.1
	50	JK	93.36	9.73	-0.05	93.4	10.06	-0.05	91.54	10.16	-0.17	91.54	10.55	-0.18	92.84	12.9	-0.05
		Boot	93.72		-0.03	93.64		-0.03	93.28		-0.09	93.28		-0.09	93		-0.06
	60	JK	94.16	7.39	-0.02	94.12	7.72	-0.02	91.97	7.79	-0.15	91.93	8.1	-0.15	93.48	9.49	0.02
		Boot	93.56		0.01	93.72		0	93.29		-0.04	93.67		-0.05	93.56		0.01
MAR	30	JK	93.04	22.9	-0.06	93.2	23.22	-0.06	91.19	23.73	-0.15	91.19	23.94	-0.14	92.32	31.82	-0.09
		Boot	92.6		-0.05	92.72		-0.05	92		-0.08	92		-0.08	91.6		-0.11
	40	JK	94.36	12.69	-0.01	94.36	12.96	-0.01	92.58	13.09	-0.11	92.46	13.3	-0.11	93.72	17.07	-0.01
		Boot	94.24		0.01	94.12		0.01	93.35		-0.03	93.35		-0.03	92.84		-0.03
	50	JK	92.56	9.62	-0.07	92.44	9.89	-0.07	90.36	9.98	-0.17	90.15	10.16	-0.17	93.08	12.74	-0.05
		Boot	92.92		-0.06	93.2		-0.06	91.94		-0.11	92.28		-0.11	92.64		-0.07
	60	JK	93.28	7.11	-0.02	93.24	7.33	-0.02	91.1	7.55	-0.15	91.14	7.74	-0.14	93.52	9.43	-0.02
		Boot	93.44		-0.01	93.44		0	92.39		-0.07	92.56		-0.07	93.32		-0.04

Table C.4: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is drawn according to a complex design and the reporting rate is 30% based on 3,000 replicates for scenarios: Large Catch Case and NMAR.

			PSA						APSA						Ratio Estimator		
Scenario	n_{PSU}		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
Large Catch Case	30	JK	28.88	34.42	-0.14	28.96	33.61	-0.14	69.61	38.09	-0.63	57.5	37.05	-0.58	94.6	11.13	-0.05
		Boot	31.44		-0.12	31.96		-0.12	77.2		-0.5	64.96		-0.46	93.6		-0.07
	40	JK	6.72	17.92	-0.12	7.32	17.28	-0.12	69.39	16.4	-0.69	41.05	17.51	-0.71	94.76	5.86	0.01
		Boot	7.48		-0.09	8.08		-0.09	79.92		-0.48	47.57		-0.62	93.44		-0.03
	50	JK	2.04	11.99	-0.1	2.48	11.51	-0.11	64.26	10.87	-0.74	24.2	12.78	-0.76	94.84	3.93	0.07
		Boot	2.04		-0.07	2.52		-0.07	77.09		-0.53	28.6		-0.68	94.12		0.02
	60	JK	0.64	9.28	-0.13	0.8	8.67	-0.12	58.3	7.8	-0.76	17.45	8.81	-0.76	95.16	3.08	0.06
		Boot	0.64		-0.08	0.84		-0.06	71.17		-0.54	20.55		-0.68	95.08		0.02
NMAR	30	JK	92.12	23.24	-0.09	92.04	23.18	-0.09	76.94	24.12	-0.59	73.09	24.65	-0.57	92.72	25.47	-0.05
		Boot	92		-0.06	91.84		-0.06	85.33		-0.35	80.97		-0.35	92.28		-0.05
	40	JK	88.88	12.1	-0.04	88.28	12.07	-0.03	68.41	13.37	-0.66	57.16	14.52	-0.65	92.48	13.95	-0.02
		Boot	89.36		0	89.2		0.01	80.6		-0.41	69.96		-0.43	92.28		-0.04
	50	JK	84.48	8.76	-0.07	83.08	8.78	-0.07	62.86	10.17	-0.71	46.17	11.57	-0.7	93.32	10.09	-0.03
		Boot	85.48		-0.02	84.44		-0.02	77.1		-0.47	59.62		-0.5	93.48		-0.04
	60	JK	83	6.53	-0.04	81.6	6.53	-0.03	61	7.6	-0.72	42.3	8.79	-0.71	93.56	7.35	0.03
		Boot	84.6		0.01	82.8		0.02	75.95		-0.46	57.25		-0.5	93.04		0.02

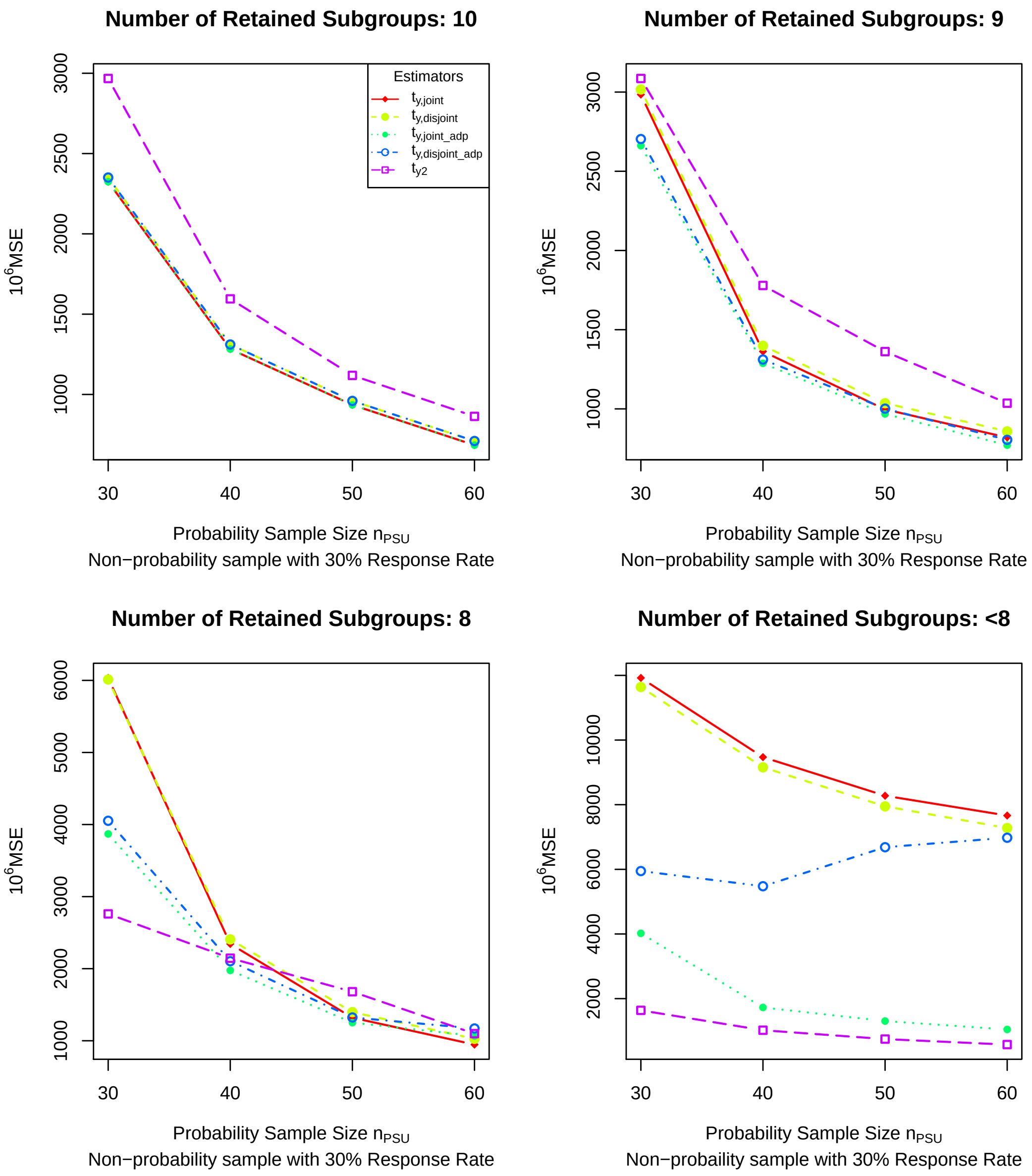


Figure C.2: Empirical MSE of the five estimators when the probability sample is according to a complex design and the reporting rate is 30% based on 3,000 replicates for different number of retained subgroups: 10, 9, 8 and Less Than 8.

Table C.5: Relative bias of the five estimators when the probability sample is drawn according to a complex design and the reporting rate is 30% based on 3,000 replicates for different number of retained subgroups: 10, 9, 8 and Less Than 8.

		PSA		APSA		Ratio Estimator
$\#subgroup$	n_{PSU}	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint-adp}$	$\hat{t}_{y,disjoint-adp}$	$\hat{t}_{y2}$
10	30	0.61	0.63	0.61	0.63	0.75
	40	0.24	0.27	0.24	0.27	0.45
	50	-0.11	-0.08	-0.11	-0.08	0.24
	60	-0.24	-0.22	-0.24	-0.22	-0.03
9	30	2.21	2.26	1.41	1.41	1.32
	40	0.04	0.11	0.02	0.05	0.32
	50	0.35	0.41	0.06	0.04	0.77
	60	-0.16	-0.1	-0.03	-0.05	0.26
8	30	9.89	9.95	6.65	7.15	2.24
	40	3.64	3.8	2.31	2.61	2.91
	50	1.34	1.45	0.88	1.03	1.22
	60	0.51	0.64	0.39	0.53	0.49
Less Than 8	30	19.69	19.56	7.86	12.26	0.21
	40	18.22	18.05	3.27	12.68	0.21
	50	17.15	16.98	1.73	14.78	0.22
	60	16.53	16.33	0.59	15.35	0.19

Table C.6: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is drawn according to a complex design and the reporting rate is 30% based on 3,000 replicates for number of retained subgroups: 10 and 9.

			PSA						APSA						Ratio Estimator		
$\#subgroup$	n_{PSU}		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
10	30	JK	93.79	23.22	-0.06	93.79	23.48	-0.06	93.79	23.22	-0.06	93.79	23.48	-0.06	92.57	29.64	-0.03
		Boot	93.41		-0.04	93.33		-0.04	93.41		-0.04	93.33		-0.04	92.3		-0.04
	40	JK	94.35	12.81	-0.01	94.15	13.09	-0.01	94.35	12.81	-0.01	94.15	13.09	-0.01	94.19	15.92	0.05
		Boot	94.15		0	94.15		0	94.15		0	94.15		0	93.87		0.04
	50	JK	92.87	9.34	-0.04	92.83	9.59	-0.04	92.87	9.34	-0.04	92.83	9.59	-0.04	93.91	11.19	0.07
		Boot	93.39		-0.03	93.87		-0.03	93.39		-0.03	93.87		-0.03	93.31		0.04
	60	JK	94.05	6.83	0.03	93.93	7.09	0.03	94.05	6.83	0.03	93.93	7.09	0.03	93.73	8.64	0.09
		Boot	93.46		0.05	93.65		0.04	93.46		0.05	93.65		0.04	93.97		0.06
9	30	JK	89.15	29.01	-0.25	89.01	29.29	-0.24	90.55	27.21	-0.31	90.38	27.62	-0.31	92.2	30.5	-0.07
		Boot	89.85		-0.27	89.78		-0.26	89.56		-0.22	89.56		-0.22	90.88		-0.12
	40	JK	91.18	13.63	-0.07	91.47	13.99	-0.07	93.02	12.88	-0.2	93.26	13.12	-0.2	92.84	17.77	-0.02
		Boot	92.52		-0.08	92.74		-0.08	93.26		-0.08	93.31		-0.07	91.32		-0.08
	50	JK	90.46	9.96	-0.03	90.26	10.35	-0.04	93.02	9.68	-0.21	92.9	10.01	-0.22	92.79	13.56	-0.04
		Boot	92.23		-0.05	92.1		-0.05	92.9		-0.11	92.84		-0.1	92.67		-0.08
	60	JK	91.46	8.7	-0.07	91.18	8.58	-0.07	93.62	8.19	-0.23	93.5	9.03	-0.23	93.79	10.35	0.01
		Boot	93.11		-0.08	93.66		-0.08	93.03		-0.13	93.15		-0.14	93.85		-0.05

Table C.7: Coverage rate and relative bias of the jackknife and bootstrap variance estimates for the five estimators with their empirical variance ($\times 10^8$) when the probability sample is drawn according to a complex design and the reporting rate is 30% based on 3,000 replicates for number of retained subgroups: 8 and Less Than 8.

			PSA						APSA						Ratio Estimator		
$\#_{subgroup}$	n_{PSU}		$\hat{t}_{y,joint}$			$\hat{t}_{y,disjoint}$			$\hat{t}_{y,joint\_adp}$			$\hat{t}_{y,disjoint\_adp}$			$\hat{t}_{y2}$		
			Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias	Coverage (%)	Var	RelBias
8	30	JK	81.13	45.12	-0.5	80.83	44.65	-0.49	83.14	32.3	-0.5	81.63	33.06	-0.51	92.1	26.71	-0.11
		Boot	81.05		-0.5	81.49		-0.49	87.4		-0.37	85.8		-0.39	91		-0.15
	40	JK	87.94	21.12	-0.38	87.94	21.54	-0.37	85.04	18.84	-0.47	84.03	19.88	-0.48	91.84	20.07	-0.11
		Boot	86.6		-0.41	86.48		-0.4	86.76		-0.35	85.47		-0.36	90.01		-0.19
	50	JK	91.35	12.85	-0.2	91.5	13.55	-0.2	85.74	12.39	-0.39	85.21	13.05	-0.39	92.41	16.48	-0.14
		Boot	91.96		-0.25	90.9		-0.25	89.08		-0.27	88.56		-0.28	91.05		-0.22
	60	JK	93.14	9.4	-0.12	93.62	10.17	-0.14	85.45	10.64	-0.42	84.73	11.65	-0.43	92.98	10.93	0.03
		Boot	92.98		-0.16	92.82		-0.18	89.82		-0.26	90.36		-0.29	91.23		-0.06
Less Than 8	30	JK	57.53	52.6	-0.56	57.82	50.68	-0.55	72.34	30.8	-0.67	63.45	33.57	-0.66	93.46	16.36	-0.06
		Boot	58.2		-0.57	58.32		-0.56	79.97		-0.52	70.13		-0.54	92.36		-0.12
	40	JK	48.28	36.75	-0.63	48.28	34.64	-0.62	69.13	15.45	-0.69	50.36	26.56	-0.73	93.37	10.22	-0.07
		Boot	48.58		-0.63	48.8		-0.62	80.07		-0.48	59.02		-0.63	92.28		-0.12
	50	JK	44.71	30.87	-0.7	44.32	28.65	-0.68	64.51	12.52	-0.76	38.05	29.51	-0.8	93.92	7.48	-0.06
		Boot	44.95		-0.69	44.66		-0.67	77.54		-0.57	46.21		-0.73	93.14		-0.11
	60	JK	43.55	28.03	-0.74	42.92	25.37	-0.72	60.62	10.38	-0.79	32.71	29.19	-0.84	94.17	5.77	-0.04
		Boot	43.82		-0.73	43.02		-0.71	73.78		-0.62	41.3		-0.77	93.25		-0.09

APPENDIX D

The Results from the PSA Method for Red Porgy

Table D.1: Numbers of trips for Red Porgy from the MRIP and ELB samples within each subgroup based on the PSA method.

Subgroup	Number of	Number of	Catch Mean		P-value
Number	MRIP Trips	ELB Trips	MRIP	ELB	
1	473	127	0.09	0.07	0.09
2	220	323	0.95	0.98	0.11
3	143	428	2.72	1.56	0.40
4	85	496	2.46	2.10	0.73
5	46	562	1.41	1.51	0.37
6	37	209	2.51	2.75	0.80
7	31	574	2.39	2.78	0.30
8	9	503	2.33	3.33	0.96
9	8	568	4.75	4.49	0.24
10	4	559	2.25	6.14	0.24

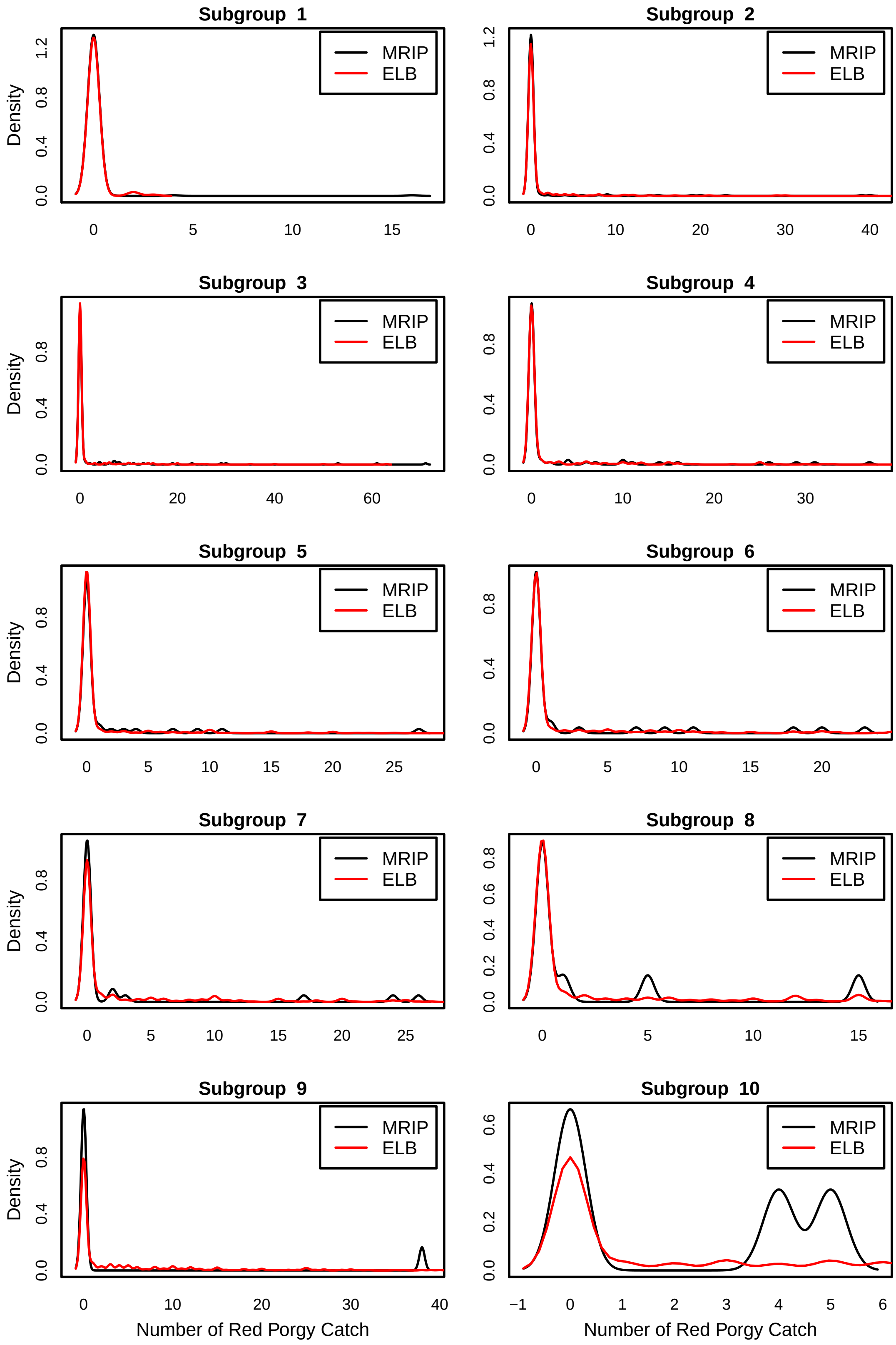


Figure D.1: Distributions of the Red Porgy catch from the MRIP and ELB samples within each subgroup based on the PSA method.

BIBLIOGRAPHY

Bingchen Liu, Lynne Stokes, Tara Topping, and Greg Stunz. 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, 2017.

Michael Robbins, Ghosh-Dastidar Bonnie, and Rajeev Ramchand. Blending of probability and convenience samples as applied to a survey of military caregivers. In *2015 Joint Statistical Mettings*, 2015.

Frederick F Stephan. History of the uses of modern sampling procedures. *Journal of the American statistical Association*, 43(241):12–39, 1948.

Martin R Frankel and Lester R Frankel. Fifty years of survey sampling in the united states. *The Public Opinion Quarterly*, 51:S127–S138, 1987.

Wayne A Fuller. *Sampling statistics*, volume 560. John Wiley & Sons, 2011.

Sharon L Lohr. *Sampling: Design and Analysis: Design and Analysis*. Chapman and Hall/CRC, 2019.

Mick P Couper. Technology trends in survey data collection. *Social Science Computer Review*, 23(4):486–501, 2005.

Reg Baker, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau. Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2):90–143, 2013.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

Sharon L Lohr, Trivellore E Raghunathan, et al. Combining survey data with other data sources. *Statistical Science*, 32(2):293–312, 2017.

Rachel A Grana, Lucy Popova, and Pamela M Ling. A longitudinal analysis of electronic cigarette use and smoking cessation. *JAMA internal medicine*, 174(5):812–813, 2014.

Matthias Schonlau, Mick P Couper, et al. Options for conducting web surveys. *Statistical Science*, 32(2):279–292, 2017.

Richard Valliant. Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 2019.

ED Le Cren. Some factors regulating the size of populations of freshwater fish: With 3 figures in the text. *Internationale Vereinigung für Theoretische und Angewandte Limnologie: Mitteilungen*, 13(1):88–105, 1965.

F Jay Breidt, Jean D Opsomer, and Chien-Min Huang. Model-assisted survey estimation with imperfectly matched auxiliary data. In *International Conference of the Thailand Econometrics Society*, pages 21–35. Springer, 2018.

KH Pollock, SC Turner, and CA Brown. Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20(2):117–124, 1994.

Ingram Olkin. Multivariate ratio estimation for finite populations. *Biometrika*, 45(1/2):154–165, 1958.

S Lynne Stokes, Benjamin M Williams, Ryan McShane, and Shalima Zalsha. The impact of nonsampling errors on estimators of catch from electronic reporting systems. *Journal of Survey Statistics and Methodology*, 2019a.

Leslie Kish. *Survey sampling*, volume 60. Wiley-Interscience, 1995.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Joseph Berkson. Are there two regressions? *Journal of the american statistical association*, 45(250):164–180, 1950.

Lynne Stokes, Ryan McShane, Ben Williams, and Shalima Zalsha. Smu recreational fisheries report for cls america. 2019b.

Sunghee Lee. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22(2):329, 2006.

Sunghee Lee and Richard Valliant. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–343, 2009.

Matthias Schonlau, Arthur Van Soest, Arie Kapteyn, and Mick Couper. Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, 37 (3):291–318, 2009.

Richard Valliant and Jill A Dever. Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137, 2011.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Yilin Chen, Pengfei Li, and Changbao Wu. Doubly robust inference with non-probability survey samples. *arXiv preprint arXiv:1805.06432*, 2018.

Jelke Bethlehem. Selection bias in web surveys. *International Statistical Review*, 78(2): 161–188, 2010.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Douglas Rivers. Sampling for web surveys. In *Joint Statistical Meetings*, 2007.

Michael R Elliott, Richard Valliant, et al. Inference for nonprobability samples. *Statistical Science*, 32(2):249–264, 2017.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

Jae Kwang Kim and Zhonglei Wang. Sampling techniques for big data analysis. *International Statistical Review*, 87:S177–S191, 2019.

Kirk Wolter. *Introduction to variance estimation*. Springer Science & Business Media, 2007.

Ramón Ferri-García and María del Mar Rueda. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS One*, 15(4):e0231500, 2020.