Discrete Ranked Set Sampling

Heng Cui
Southern Methodist University, hcui@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds
Part of the Statistical Methodology Commons

Recommended Citation
https://scholar.smu.edu/hum_sci_statisticalscience_etds/2

This Dissertation is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
DISCRETE RANKED SET SAMPLING

Approved by:

________________________________________
Dr. Lynne Stokes
Committee Chair

________________________________________
Dr. Min Chen

________________________________________
Dr. Hon Keung Tony Ng

________________________________________
Dr. Cornelis Jr. Potgieter

________________________________________
Dr. Xinlei Wang
DISCRETE RANKED SET SAMPLING

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 Statistical Science
by
Heng Cui

B.S., Information and Computational Science, Shandong University, China
M.S., Financial Mathematics, Shandong University, China

May 7, 2018
ACKNOWLEDGMENTS

Completion of this dissertation is impossible without the supports of several people. First of all, I would like to thank my advisor, Dr. Lynne Stokes, for her encouragement, guidance, and supports through the last two years. Dr. Stokes has provided me so many insightful suggestions and advice to help me finish my dissertation. She also made me realize the importance of paying attention to detail. Her advice and careful editing improved this dissertation to another level.

I am extremely grateful for all valuable suggestions and concise comments from the other committee members- Dr. Tony, Dr. Potgieter, Dr. Sherry and Dr. Chen. Thank for your time and academic efforts to help me improve my dissertation. To Dr. Tony, Dr. Potgieter and Dr. Sherry, I was so lucky to take classes from you. Your expertise in Statistics has inspired me and helped me learn about the beauty of Statistics.

I would also like to express my thanks to all of my friends. Thank you for listening to me, offering me advice and supporting me throughout the journey.

Finally, I would like to acknowledge with gratitude, the support and love of my family- my parents, my brother and my wife. Thank you all for encouraging me in all of my pursuits and inspiring me to follow my dreams. You are the most important people to me in the world and I dedicate this dissertation to you.
Ranked set sampling (RSS) is an efficient data collection framework compared to simple random sampling (SRS). It is widely used in various application areas such as agriculture, environment, sociology, and medicine, especially in situations where measurement is expensive but ranking is less costly. Most past research in RSS focused on situations where the underlying distribution is continuous. However, it is not unusual to have a discrete data generation mechanism. Estimating statistical functionals are challenging as ties may truly exist in discrete RSS. In this thesis, we started with estimating the cumulative distribution function (CDF) in discrete RSS. We proposed two methods to incorporate the information brought by ties. The first method is based on the idea of Frey (2012), which only works for the balanced RSS. The second one is based on the NPMLE method proposed by Kvam and Samaniego (1994). The second method can be applied in both balanced and unbalanced RSS. By simulation studies, we showed that the new methods improve the efficiency of estimation. Later, we proposed the corresponding plug-in estimators for the population mean and the population variance. The new estimators showed higher efficiency compared to the existing estimators in the literature.

Another problem considered in this thesis is to improve the estimation efficiency of each order stratum CDF when tie information is not available. We proposed a new estimator by imposing uniformly stochastic ordering constraint on the order strata CDF’s. By using the "ranking" relationship between the order strata CDF’s, the new estimator showed a higher efficiency for the strata except the edge strata (the smallest and largest order stratum).
TABLE OF CONTENTS

LIST OF FIGURES ................................................................. viii
LIST OF TABLES ................................................................. x

CHAPTER

1. INTRODUCTION ............................................................... 1
   1.1. Research Background .................................................. 1
   1.2. Discrete Ranked Set Sampling Procedures and Variations .......... 2
   1.3. Research Objectives ................................................... 4

2. ESTIMATORS OF THE CUMULATIVE DISTRIBUTION FUNCTION BY
   USING RSS ................................................................. 5
   2.1. Empirical Distribution Function ...................................... 5
   2.2. Alternative to the EDF: Frey’s Estimator .......................... 7
   2.3. Nonparametric Maximum Likelihood Estimator of the CDF ....... 11
   2.4. Resampling Methods for Ranked Set Sampling .................... 16
   2.5. Simulation Study ..................................................... 19
   2.6. Conclusion .......................................................... 23

3. ON ESTIMATING THE POPULATION MEAN ............................. 24
   3.1. Background .......................................................... 24
   3.2. Frey’s Estimator of the Population Mean .......................... 26
   3.3. Nonparametric Maximum Likelihood Estimator of \( \mu \) ............ 27
   3.4. Simulation Study ..................................................... 28
      3.4.1. Performance of Estimators Under Balanced Discrete Ranked Set
              Sampling ......................................................... 28
      3.4.2. Performance of Estimators Under Unbalanced Discrete Ranked
              Set Sampling ..................................................... 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.3</td>
<td>Inference on Population Mean</td>
<td>35</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusion</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>ON ESTIMATING THE POPULATION VARIANCE</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Background</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>New Estimators of Population Variance</td>
<td>41</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation Study</td>
<td>42</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Comparison of Variance Estimators</td>
<td>42</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Inference on Population Variance</td>
<td>47</td>
</tr>
<tr>
<td>4.4</td>
<td>Conclusion</td>
<td>53</td>
</tr>
<tr>
<td>5</td>
<td>ON ESTIMATING THE ORDER STRATUM CDF BY USING DISCRETE RANKED SET SAMPLING</td>
<td>54</td>
</tr>
<tr>
<td>5.1</td>
<td>Literature Review</td>
<td>54</td>
</tr>
<tr>
<td>5.2</td>
<td>Three Types of Stochastic Ordering</td>
<td>55</td>
</tr>
<tr>
<td>5.3</td>
<td>Estimators under Ordinary Stochastic Ordering</td>
<td>57</td>
</tr>
<tr>
<td>5.4</td>
<td>Estimating Under Uniformly Stochastic Ordering</td>
<td>58</td>
</tr>
<tr>
<td>5.5</td>
<td>Simulation Studies</td>
<td>59</td>
</tr>
<tr>
<td>5.6</td>
<td>Conclusions</td>
<td>63</td>
</tr>
<tr>
<td>6</td>
<td>DISCUSSION AND FUTURE DIRECTIONS</td>
<td>64</td>
</tr>
<tr>
<td>6.1</td>
<td>Ranking Error Models</td>
<td>64</td>
</tr>
<tr>
<td>6.2</td>
<td>Variations of Ranked Set Sampling</td>
<td>65</td>
</tr>
</tbody>
</table>

APPENDIX

| A.   | DISCRETE ORDER STATISTICS                                             | 66   |
| BIBLIOGRAPHY |                                                                 | 72   |
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>RE of $\hat{p}_j^c$ for $H = 2, 3, 4, \text{ and } 5.$</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Original samples and corresponding weights</td>
<td>8</td>
</tr>
<tr>
<td>2.3</td>
<td>Relative efficiency of CDF estimators for discrete uniform distributions</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>Relative efficiency of CDF estimators for binomial distributions</td>
<td>21</td>
</tr>
<tr>
<td>2.5</td>
<td>Relative efficiency of CDF estimators for Poisson distributions</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Relative efficiency of mean estimators for discrete uniform distributions ($N_0$)</td>
<td>29</td>
</tr>
<tr>
<td>3.2</td>
<td>Relative efficiency for binomial distributions ($n = 10, p$)</td>
<td>30</td>
</tr>
<tr>
<td>3.3</td>
<td>Relative efficiency of mean estimators for Poisson distributions ($\lambda$)</td>
<td>31</td>
</tr>
<tr>
<td>3.4</td>
<td>Relative efficiency of mean estimators versus sampling proportion from 1st order stratum for discrete uniform ($N_0 = 5, 10$), Poisson ($\lambda = 1, 3$) and binomial ($n = 10, p = .2, .5$) for $H = 2.$</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Relative efficiency of population variance estimators for discrete uniform distributions ($N_0$).</td>
<td>44</td>
</tr>
<tr>
<td>4.2</td>
<td>Relative efficiency of population variance estimators for binomial distributions ($n = 10, p$).</td>
<td>45</td>
</tr>
<tr>
<td>4.3</td>
<td>Relative efficiency for Poisson distributions ($\lambda$).</td>
<td>46</td>
</tr>
<tr>
<td>4.4</td>
<td>Histogram of $\hat{\sigma}_{NPMLE}^2$ when ranked set samples are generated from binomial ($10, 0.5$). The set size is 5 and the cycle size is 5.</td>
<td>48</td>
</tr>
<tr>
<td>4.5</td>
<td>Coverage probability of confidence intervals produced by the bootstrap procedure when ranked set samples or simple random samples generated for binomial ($10, 0.5$) when $H = 5$. For simple random sample, the $x$-axis denotes equivalent cycle size; that is the sample size divided by $H$.</td>
<td>51</td>
</tr>
<tr>
<td>5.1</td>
<td>Relative efficiency of $F_{i}^{\text{US}}$ ($i = \text{iso, uso}$) for discrete uniform distribution ($N_0$). The number $r$ on each line respresents $r$-th order stratum.</td>
<td>61</td>
</tr>
</tbody>
</table>
5.2 Relative efficiency of $\hat{F}_{[p]}^i (i = iso, uso)$ for binomial($n = 10, p$) distribution. The number $r$ on each line represents $r$-th order stratum. ................. 62

5.3 Relative efficiency of $\hat{F}_{[r]}^i (i = iso, uso)$ for Poisson distribution ($\lambda$). The number $r$ on each line represents $r$-th order stratum. ...................... 62
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>35</td>
</tr>
<tr>
<td>3.3</td>
<td>38</td>
</tr>
<tr>
<td>4.1</td>
<td>52</td>
</tr>
<tr>
<td>A.1</td>
<td>70</td>
</tr>
</tbody>
</table>
This is dedicated to my family- my parents, my brother and my wife.
1.1. Research Background

Ranked set sampling (RSS) is an efficient data collection framework compared to simple random sampling (SRS). It is widely used in various application areas such as agriculture, environment, sociology, and medicine, especially in situations where measurement is expensive but ranking is less costly.

RSS was first proposed by McIntyre (1952) for estimating mean pasture yields. It improved the efficiency of estimating the population mean by incorporating the investigators’ judgment. Takahasi and Wakimoto (1968) provided the theoretical foundations of RSS. They showed that the relative efficiency of the mean estimator by using RSS compared to SRS is bounded and the upper bound is achieved when the underlying distribution is a uniform distribution. Later, the research on ranked set sampling developed in two major directions. One direction of research on RSS was to estimate different parameters other than the mean, including the population variance, cumulative distribution function (CDF), quantile function, etc. For example, Stokes and Sager (1988) proposed the empirical distribution function (EDF) based on a ranked set sample. Kvam and Samaniego (1994) provided the nonparametric maximum likelihood estimator of the CDF, which was shown to be more precise compared to the EDF when no ranking error exist. Another direction of research was to consider the variations of the RSS design. One of the most important variations was judgement post-stratification (JPS) proposed by MacEachern et al. (2004). JPS attracted a lot of research because of its practical flexibilities.

However, most past research in RSS or JPS focused on situations where the underlying distribution is continuous. However, it is not unusual to have a discrete data generation
mechanism. The only available literature considering the RSS with discrete underlying distributions is Barabesi and Pisani (2002). One important difference between discrete cases and continuous cases is that ties may truly exist in discrete cases. Frey (2012) discussed the perceived tie problem in continuous JPS. A perceived tie is declared if the ranker is not sure about the ranking between two units. Frey (2012) showed that using information about perceived ties can improve the relative efficiency of mean estimators. Later in this thesis, we will show their proposed estimator can be adapted into discrete RSS, where the perceived ties are actual ties.

1.2. Discrete Ranked Set Sampling Procedures and Variations

Generally, the process of selecting a ranked set sample can be briefly summarized as following:

First, draw a set of $H$ sample units randomly from the population and then rank those units visually or by some inexpensive method. $H$ is called the set size and usually set to be small (less than 10) in order to avoid ranking errors. The unit judged smallest is then chosen to be measured in an accurate way which may be expensive or time-consuming. Then repeat the same procedure until we get $n_1$ units judged smallest to measure.

Similarly, we can sample $n_r$ units judged $r$-th ($r = 2, \cdots, H$) smallest to measure. After measuring those units accurately, we get a ranked set sample. Let $\{X_{[r]}[i]; r = 1, \cdots, H; i = 1, \cdots, n_r\}$ denote the ranked set sample, where $X_{[r]}[i]$ is the measured value of $i$-th unit with rank $r$. If $n_1 = n_2 = \cdots = n_H = n$, the design is balanced and $n$ is called cycle size.

An unbalanced ranked set sample differs from a balanced one in that the measured order statistics no longer must be included an equal number of times in the full sample, as long as units from each order stratum are included at least once. In the following discussion, let $N$ denote the sample size of a ranked set sample, i.e. $N = \sum_{r=1}^{H} n_r$.

The whole process is analogous to stratified sampling. Each order statistic is analogous to one stratum in stratified sampling. Therefore, a ranked set sample with set size $H$ is similar to a stratified sample with $H$ strata.
An important variation of ranked set sampling is judgment post-stratification (JPS) proposed by MacEachern et al. (2004). In JPS, a simple random sample of size \( N \) is selected and measured first. Then \( N \) sets of samples of size \( H - 1 \) are selected to work as the comparison samples. JPS is similar to unbalanced RSS except that the sample size from each stratum is determined randomly by using a multinomial distribution, i.e. \( (n_1, \ldots, n_H) \sim \text{Multinomial}(N; \frac{1}{H}, \ldots, \frac{1}{H}) \). Although JPS was shown to be less efficient than RSS, it still attracted a lot of research attention because of its practical usefulness. Besides JPS, there were some other useful variations of RSS in the literature; see Al-Saleh and Al-Kadiri (2000), Al-Saleh and Al-Omari (2002), Hossain and Muttlak (1999), etc.

The main focus in this thesis is to consider the applications of RSS and its variations when the underlying distribution is discrete. So it is important to make clear the differences between discrete RSS and continuous RSS first. The main difference is in the ranking process. In continuous RSS, the rankers should always assign different ranks to any two units if their ability to rank is perfect. This is not the case in the discrete distribution. A perfect ranker should claim a tie when they believe there is no difference between two units. To get the units for measurement, the ranker randomly selects one unit from the tied units. Then the number of ties can be viewed as extra information, and therefore it seems plausible that inference may be improved by using tie information. Unlike discrete RSS, ties are declared in continuous RSS because the rankers are unsure about the true ranks of some units. In other words, claiming a tie between two units in continuous RSS is actually making ranking errors in some sense.

Another difference between discrete RSS and continuous RSS is related to the order statistics theory behind them. In Appendix A, we have a detailed discussion about the differences between the discrete order statistics theory and the continuous order statistics theory.

In the following discussion, we write a discrete ranked-set sample as \( \{(X_{[r]i}, l_{[r]i}, t_{[r]i}); r = 1, \ldots, H; i = 1, \ldots, n_r\} \), where \( l_{[r]i} \) and \( t_{[r]i} \) are the number of units which are judged to be less than the measured unit and the number of units which are judged equal to the measured
unit (including the measured one itself) within the set from which $X_{[r]}$ is selected. In this thesis, we focus on situations where no ranking error exist.

1.3. Research Objectives

The structure of this thesis is by topic. We first focus on the CDF estimation under discrete RSS. Our main focus is to make use of the information brought by ties. We propose two new estimators of the CDF which incorporate tie information in Chapter 2. One of them is motivated from the idea of Frey (2012). The other is an application of Kvam and Samaniego (1994) to discrete cases with moderate modifications. In Chapter 3, we propose the corresponding plug-in estimators of mean $\mu$ based on the proposed CDF estimators. The plug-in estimators are compared to those existing estimators in the literature via simulation studies. In Chapter 4, we propose a plug-in estimator of the population variance based on the NPMLE of the CDF. The plug-in estimator is compared to Stokes’s estimator in Stokes (1980) and the unbiased estimator in MacEachern et al. (2002) via simulation studies. In Chapter 5, we discuss the estimation of the CDF of each order stratum. By imposing the uniformly stochastic ordering constraint, we obtain a new estimator for each order stratum CDF. In Chapter 6, we give some ideas for future research.
CHAPTER 2
ESTIMATORS OF THE CUMULATIVE DISTRIBUTION FUNCTION BY USING RSS

2.1. Empirical Distribution Function

Estimating the cumulative distribution function (CDF) of a continuous random variable by using RSS has a long history. The most common and convenient estimator is the empirical distribution function (EDF), proposed by Stokes and Sager (1988) in the context of continuous balanced RSS. Let $X$ be a random variable with probability density (or mass) function $f(x)$ and cumulative density function $F(x)$. Let $X_{[r]}$ be $r$-th order statistic among $H$ units with probability density (or mass) function $f_{[r]}(x)$ and cumulative density function $F_{[r]}(x)$. Then the EDF, denoted by $\hat{F}^e$, is given by

$$\hat{F}^e(x) = \frac{1}{H} \sum_{r=1}^{H} \hat{F}^e_{[r]}(x),$$

(2.1)

where $\hat{F}^e_{[r]}(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_{[r]}^i \leq x\}$ is the EDF of $r$-th order statistic and $I\{\cdot\}$ is the indicator function.

Several useful propositions for the EDF from a balanced ranked set sample were provided by Stokes and Sager (1988):

1. $\hat{F}^e$ is an unbiased and consistent estimator of $F$.

2. $Var(\hat{F}^e) = \frac{1}{nH^2} \sum_r F_{[r]}(1 - F_{[r]})$.

3. For a fixed $H$ and $x$, $\frac{\hat{F}^e(x) - F(x)}{\sqrt{Var(\hat{F}^e(x))}}$ converges to a standard normal distribution as $n \to \infty$.

Although all three propositions were given in the context of continuous RSS, they also hold for discrete RSS. Moreover, the EDF with a discrete ranked set sample is equivalent to
an empirical estimator of pmf. Denote the support of a discrete random variable $X$ as $S$. Let $\hat{p}_{rj}^e$ and $\hat{p}_r^e$ denote the empirical estimator of $p_{rj} = f_{[r]}(x_j)$ and $p_j = f(x_j)$, where $x_j \in S$. $\hat{p}_{rj}^e$ and $\hat{p}_r^e$ are given by

\[
\hat{p}_{rj}^e = \frac{1}{n} \sum_{i=1}^{n} I\{X_{[r]}i=x_j\}, \forall x_j \in S
\]

\[
\hat{p}_r^e = \frac{1}{H} \sum_{r=1}^{H} \hat{p}_{rj}^e.
\]

Then the previous propositions can be rewritten in terms of $\hat{p}_{rj}^e$ and $\hat{p}_r^e$:

1. $\hat{p}_{rj}^e$ is an unbiased estimator of $p_{rj}$ and, therefore, $\hat{p}_r^e$ is an unbiased estimator of $p_j$.
2. $Var(\hat{p}_{rj}^e) = \frac{1}{n} p_{rj}(1 - p_{rj})$ and $Var(\hat{p}_r^e) = \frac{1}{nH} \sum_r p_{rj}(1 - p_{rj})$.
3. $\frac{\hat{p}_{rj}^e - p_{rj}}{Var(\hat{p}_r^e)}$ converges to a standard normal distribution as $n \to \infty$.

Define the relative efficiency (RE) of $\hat{p}_r^e$ as the ratio of mean square errors between the SRS estimator of $p_j$ from a simple random sample of size $nH$, denoted by $\hat{p}_j^{SRS}$, and the RSS estimator $\hat{p}_r^e$. RE is equivalent to the ratio of the variance of the two estimators because both of them are unbiased estimators of $p_j$, which is

\[
RE = \frac{Var(\hat{p}_j^{SRS})}{Var(\hat{p}_r^e)} = \frac{p_j(1 - p_j)}{\frac{1}{H} \sum_r p_{rj}(1 - p_{rj})}.
\]

$RE$ of $\hat{p}_r^e$ is the same as the relative precision of the RSS estimator of the population proportion in Chen et al. (2006). In Figure 2.1, we show the relative efficiency of $\hat{p}_r^e$ for $H = 2, 3, 4$, and 5.
Figure 2.1. RE of $\hat{p}_j$ for $H = 2, 3, 4,$ and $5.$

From Figure 2.1, it appears that RE is bounded by 1 and $(H + 1)/2$, the same as the bounds of the RE of the mean estimator under continuous balanced RSS. The rigorous proof of this conjecture, however, is still an open question. Moreover, unlike continuous RSS, the upper bound of RE is no longer always achieved. More properties of $RE$ of $\hat{p}_j$, including the limiting behavior, the symmetry property etc, were given in Chen et al. (2006). For more details, see Chen et al. (2006).

2.2. Alternative to the EDF: Frey’s Estimator

In this section, we introduce an alternative to the EDF for discrete random variables, which incorporates tie information in an ad-hoc way. This estimator stems from an idea of Frey (2012) and only works for the balanced discrete RSS. The resulting estimator of the $r$-th
stratum CDF and the population CDF are denoted by \( \hat{F}_{[r]}(x) \) and \( \hat{F}_{\text{Frey}}(x) \), respectively. The idea is to assign equal weights to all observed tied values within each ranked set. This idea is illustrated by the following example.

**Example 2.1** Suppose we have 6 sets of samples of size 3 from a discrete distribution (i.e. \( H = 3, n = 2 \)), as shown in Figure 2.2. Following the procedures in Chapter 1.2, we get the accurately measured value \((1, 2, 3, 3, 3, 4)\) and the number of observed ties for each measured unit (including itself) \((2, 1, 3, 1, 1, 1)\). Then we assign equal weights to all units which are tied to the measured value within every set. For those units which are not measured or claimed tied to the measured units, we assign weight 0 to them. The weight for each unit is shown in Figure 2.2.

Because information about \( r \)-th order statistic is also available from the samples in which the \( s \)-th (\( s \neq r \)) order statistic is measured, due to the knowledge of ties, we have a larger
sample size than \( n = 2 \) for estimating \( F_{[r]} \). Frey’s method uses tie information as follows:

\[
\hat{F}^{\text{Frey}}_{[1]}(x) = \begin{cases} 
0, & x < 1 \\
\frac{1/2}{1/2 + 1 + 1/3}, & 1 \leq x < 2 \\
\frac{1/2}{1/2 + 1 + 1/3} + \frac{1}{1/2 + 1 + 1/3} = \frac{9}{11}, & 2 \leq x < 3 \\
1, & x \geq 3
\end{cases}
\]

(2.1)

\[
\hat{F}^{\text{Frey}}_{[2]}(x) = \begin{cases} 
0, & x < 1 \\
\frac{1/2}{1/2 + 1 + 1/3} = \frac{3}{11}, & 1 \leq x < 3 \\
1, & x \geq 3
\end{cases}
\]

(2.2)

\[
\hat{F}^{\text{Frey}}_{[3]}(x) = \begin{cases} 
0, & x < 3 \\
\frac{1/3 + 1}{1/3 + 1 + 1} = \frac{4}{7}, & 3 \leq x < 4 \\
1, & x \geq 4
\end{cases}
\]

(2.3)

By using the relation \( F(x) = \frac{1}{H} \sum_{r=1}^{H} F_{[r]}(x) \), we can construct an estimate of the overall CDF:

\[
\hat{F}^{\text{Frey}}(x) = \begin{cases} 
0, & x < 1 \\
\frac{2}{11}, & 1 \leq x < 2 \\
\frac{4}{11}, & 2 \leq x < 3 \\
\frac{6}{7}, & 3 \leq x < 4 \\
1, & x \geq 4
\end{cases}
\]

(2.3)

The estimates in (2.2) and (2.3) are equivalent to the estimates of pmf of each order stratum and the overall population, given by

9
\(\hat{f}_{\text{Frey}}(x) = \begin{cases} 
\frac{1}{2} & x = 1 \\
\frac{1}{2} + \frac{1}{1 + 1/3} = \frac{3}{11}, x = 2 \\
\frac{1}{2} + \frac{1}{1 + 1/3} = \frac{2}{11}, x = 3 \\
0, \text{ elsewhere}, 
\end{cases} \)

\(\hat{f}_{\text{Frey}}(x) = \begin{cases} 
\frac{1}{2} & x = 1 \\
\frac{1}{2} + \frac{1}{1 + 1/3} = \frac{3}{11}, x = 2 \\
\frac{1}{2} + \frac{1}{1 + 1/3} = \frac{8}{11}, x = 3 \\
0, \text{ elsewhere}, 
\end{cases} \)

\(\hat{f}_{\text{Frey}}(x) = \begin{cases} 
\frac{1}{3} + \frac{1}{1 + 1/3} = \frac{4}{7}, x = 3 \\
\frac{1}{3} + \frac{1}{1 + 1/3} = \frac{3}{7}, x = 4 \\
0, \text{ elsewhere}, 
\end{cases} \)

Now we formalize the procedure for a balanced ranked set sample. We denote a balanced ranked set sample by \(\{(X_{[r]i}, l_{[r]i}, t_{[r]i}), r = 1, \ldots, H; i = 1, \ldots, n\}\), where \(l_{[r]i}\) is the number of units which are judged to be less than the measured unit and \(t_{[r]i}\) is the number of units which are judged tied to the measured unit (including the measured one itself) within the ranked set from which \(X_{[r]i}\) is selected for measurement. For example, the sample in Example 2.1 can be written as \(\{(1, 0, 2), (2, 0, 1), (3, 0, 3), (3, 1, 1), (3, 2, 1), (4, 2, 1)\}\). Let \(x_1 < x_2 < \cdots < x_k\) denote the \(k\) distinct measured values from the ranked set sample. We define \(S_r\) to be the indices \((h, i)\) of the observations \(X_{[h]i}\) which provide information about
F_{[r]}$, either directly because $h = r$, or indirectly, because $X_{[h]}$ is tied with a unit having rank $r$. Then $S_r = \{(h, i) | l_{[h]} < r \leq l_{[h]} + t_{[h]} \}, r = 1, \cdots, H$. In Example 2.1, we have $S_1 = \{(1, 1), (1, 2), (2, 1)\}$, $S_2 = \{(1, 1), (2, 1), (2, 2)\}$, and $S_3 = \{(2, 1), (3, 1), (3, 2)\}$. Then for $r$-th stratum, the estimator of the CDF is given by

$$\hat{F}^{Frey}_{[r]}(x) = \sum_{x_{[j]} \leq x} \hat{f}^{Frey}_{[r]}(x_{[j]}), \quad (2.5)$$

where $\hat{f}^{Frey}_{[r]}(x)$ is the estimator of $f_{[r]}(x)$

$$\hat{f}^{Frey}_{[r]}(x) = \frac{\sum_{(h, i) \in S_r} \frac{I_{X_{[h]}=x}}{t_{[h]}}}{\sum_{(h, i) \in S_r} \frac{1}{t_{[h]}}} = \hat{f}^{Frey}_{[r]}(x). \quad (2.6)$$

By using the relations $F(x) = \frac{1}{H} \sum_{r=1}^{H} F_{[r]}(x)$ and $f(x) = \frac{1}{H} \sum_{r=1}^{H} f_{[r]}(x)$, we can construct the corresponding estimators for the overall population CDF and pmf. Although the estimators are not unbiased, they still outperform the empirical estimators in terms of mean integrated squared error (MISE), which will be shown in Chapter 2.5. One explanation was given by Liu (2016) who showed that Frey’s estimator is a modified version of Horvitz-Thompson estimator.

### 2.3. Nonparametric Maximum Likelihood Estimator of the CDF

Kvam and Samaniego (1994) proposed the nonparametric maximum likelihood estimator (NPMLE) of the CDF when sample units were generated from a continuous underlying distribution. Huang (1997) proved that the NPMLE is a strongly consistent estimator of the CDF. The idea of Kvam and Samaniego (1994) can be applied in discrete cases with moderate modifications, which we will provide later.

First, we will find the likelihood function for a ranked set sample from a discrete distribution, where ties of measured observations are accurately identified. Then we find the MLE of the CDF from the likelihood function. Let $x_1 < x_2 < \cdots < x_k$ denote the $k$ distinct
values measured in the ranked set sample. For any \( x_j \), define \( l_j, t_j, v_j \) as

\[
\begin{align*}
    l_j &= \sum_{r=1}^{H} \sum_{i=1}^{n_r} I_{\{X_{r|i}=x_j\}} \\
    t_j &= \sum_{r=1}^{H} \sum_{i=1}^{n_r} t_{r|i} I_{\{X_{r|i}=x_j\}} \\
    v_j &= \sum_{r=1}^{H} \sum_{i=1}^{n_r} I_{\{X_{r|i}=x_j\}}.
\end{align*}
\]

(2.7)

\( l_j, t_j, v_j \) are the number of units which are judged less than, tied with, or equal to \( x_j \) in the samples (including both comparison samples and measured samples). By (A.5), the likelihood function of the ranked set sample is given by

\[
L = \prod_{j=1}^{k} C_j F^{l_j}(x_j-)(1 - F(x_j))^{s_j}(F(x_j) - F(x_j-))^{t_j}
\]

(2.8)

where \( F(x_j-) = P(X < x_j) \), \( s_j = H * v_j - l_j - t_j \), and \( C_j \) is a constant.

Now we find an estimator \( \hat{F} \) that is restricted to the set

\[
\mathcal{F} = \{ \hat{F} | \hat{F}(x_{j-}) = \hat{F}(x_{j-1}), j = 2, ..., k \},
\]

and maximizes \( L \). \( \mathcal{F} \) excludes those estimators which assign positive probability to values which are between \( x_{j-1} \) and \( x_j \). Therefore (2.8) is equivalent to

\[
L = \prod_{j=1}^{k} C_j F^{l_j}(x_{j-1})(1 - F(x_j))^{s_j}(F(x_j) - F(x_{j-1}))^{t_j}.
\]

(2.9)

For notation simplification, let \( \phi_j \) be

\[
\begin{align*}
    \phi_0 &= P(X < x_1) \\
    \phi_j &= F(x_j), \forall j = 1, ..., k
\end{align*}
\]

(2.10)
Then the log-likelihood of the ranked set sample is

$$\ell = \sum_{j=1}^{k} (\log C_j + l_j \log \phi_{j-1} + s_j \log (1 - \phi_j) + t_j (\phi_j - \phi_{j-1}))$$  \hspace{1cm} (2.11)

Now, the problem is to find $\phi_0^*, \phi_1^*, ..., \phi_k^*$ which maximizes the log-likelihood function.

The maximum likelihood estimates of $\phi_0$ and $\phi_k$ can be easily obtained in some special cases. If $l_1 = 0$, we will set $\hat{\phi}_0 = 0$. $l_1 = 0$ corresponds to the situations that the smallest value $x_1$ is measured as or judged tied to the first order statistic within the sets from which they are sampled. Similarly, if $s_k = 0$, we will set $\hat{\phi}_k = 1$. $s_k = 0$ corresponds to the situations where the largest value $x_k$ are measured as or judged tied to the largest order statistic within the sets from which they are sampled. In the following discussion, we first assume $l_1 \neq 0$ and $s_k \neq 0$.

By taking the first derivatives w.r.t $\phi_0, \phi_1, ..., \phi_k$ and setting those functions to 0, we have

$$\frac{\partial \ell}{\partial \phi_0} = \frac{l_1}{\phi_0} - \frac{t_1}{\phi_1 - \phi_0} = 0$$

$$\frac{\partial \ell}{\partial \phi_1} = -\frac{s_1}{1 - \phi_1} + \frac{l_1}{\phi_1 - \phi_0} + \frac{l_2}{\phi_1} - \frac{t_2}{\phi_2 - \phi_1} = 0$$

$$\vdots$$

$$\frac{\partial \ell}{\partial \phi_j} = -\frac{s_j}{1 - \phi_j} + \frac{l_j}{\phi_j - \phi_{j-1}} + \frac{l_{j+1}}{\phi_j} - \frac{t_{j+1}}{\phi_{j+1} - \phi_j} = 0$$

$$\vdots$$

$$\frac{\partial \ell}{\partial \phi_k} = -\frac{s_k}{1 - \phi_k} + \frac{l_k}{\phi_k - \phi_{k-1}} = 0$$

Solving those equations explicitly is not practically feasible in most cases. An iterative algorithm was proposed by Kvam and Samaniego (1994). From the first $k$ likelihood equations,
we have

\[
\begin{aligned}
\phi_1 &= \phi_0 + t_1 \frac{\phi_0}{\Pi_1} \\
\phi_2 &= \phi_1 + t_2(-\frac{s_1}{1-\phi_1} + \frac{t_1}{\phi_1-\phi_0} + \frac{t_k}{\phi_1})^{-1}
\end{aligned}
\]

\[
\vdots
\]

\[
\begin{aligned}
\phi_{j+1} &= \phi_j + t_{j+1}(-\frac{s_{j+1}}{1-\phi_j} + \frac{t_j}{\phi_{j}-\phi_{j-1}} + \frac{t_{j+1}}{\phi_j})^{-1}
\end{aligned}
\]

\[
\vdots
\]

\[
\begin{aligned}
\phi_k &= \phi_{k-1} + t_k(-\frac{s_{k-1}}{1-\phi_{k-1}} + \frac{t_{k-1}}{\phi_{k-1}-\phi_{k-2}} + \frac{t_k}{\phi_{k-1}})^{-1}.
\end{aligned}
\]

(2.13)

Let \( H(\phi_k, \phi_{k-1}) \) be defined as

\[
H(\phi_k, \phi_{k-1}) = -\frac{s_k}{1-\phi_k} + \frac{t_k}{\phi_k - \phi_{k-1}}.
\]

From the last equation in (2.12), we have \( H(\phi_k^*, \phi_{k-1}^*) = 0 \), i.e. the maximum likelihood estimate of \( \phi_{k-1} \) and \( \phi_k \) should satisfy \( H(\phi_k^*, \phi_{k-1}^*) = 0 \). Later, we will show that \( H(\cdot, \cdot) \) can be used to determine the convergence of the algorithm.

Now we prove two theorems that guarantee the existence and uniqueness of the solution, and that the computational algorithm outlined above will find that solution. Those proofs are similar to those of Kvam and Samaniego (1994), who proved similar results for their algorithm for computing the NPMLE of the CDF in the continuous cases.

**Theorem 2.1** Let \( \phi_0^*, \phi_1^*, ..., \phi_k^* \) be the solution of the log-likelihood equations in (2.12). Then \( \phi_0^*, \phi_1^*, ..., \phi_k^* \) is the unique nonparametric MLE.
Proof: The Hessian matrix of log-likelihood function is a tridiagonal matrix, i.e.

\[
H = \begin{bmatrix}
h_{00} & h_{01} \\
h_{10} & h_{11} & h_{12} \\
& \ddots & \ddots \\
h_{k-1,1} & \cdots & h_{k-1,k} \\
h_{kk}
\end{bmatrix}
\]

where

\[
h_{j,j-1} = \frac{\partial^2 \ell}{\partial \phi_j \partial \phi_{j-1}} = \frac{t_j}{(\phi_j - \phi_{j-1})^2}, \quad j = 1, \ldots, k
\]

\[
h_{jj} = \frac{\partial^2 \ell}{\partial \phi_j^2} = \begin{cases} 
-\frac{l_j}{\phi_j^2}, & j = 0 \\
-\frac{s_j}{(1-\phi_j)^2} - \frac{t_j}{(\phi_j - \phi_{j-1})^2} - \frac{l_{j+1}}{\phi_j^2} - \frac{t_{j+1}}{(\phi_{j+1} - \phi_j)^2}, & j = 1, \ldots, k
\end{cases}
\]

\[
h_{j,j+1} = \frac{\partial^2 \ell}{\partial \phi_j \partial \phi_{j+1}} = \frac{t_{j+1}}{(\phi_{j+1} - \phi_j)^2}, \quad j = 0, \ldots, k - 1
\]

This matrix can be shown to be negative semidefinite. For any vector \( \mathbf{y} = (y_0, \ldots, y_{k+1})^\top \in \mathbb{R}^{k+1} \), we have

\[
\mathbf{y}^\top H \mathbf{y} = \sum_{j=0}^{k} h_{jj}y_j^2 + \sum_{j=1}^{k} h_{j,j-1}y_jy_{j-1} + \sum_{j=0}^{k-1} h_{j,j+1}y_jy_{j+1}
\]

\[
= -\frac{l_1}{\phi_0^2} y_0^2 - \sum_{j=1}^{k} \left( \frac{s_j}{(1-\phi_j)^2} + \frac{l_{j+1}}{\phi_j^2} \right) y_j^2 - \sum_{j=1}^{k} h_{j,j-1}y_j^2
\]

\[
- \sum_{j=1}^{k} h_{j,j+1}y_j^2 + \sum_{j=1}^{k} h_{j,j-1}y_jy_{j-1} + \sum_{j=0}^{k-1} h_{j,j+1}y_jy_{j+1}
\]

\[
= -\frac{l_1}{\phi_0^2} y_0^2 - \sum_{j=1}^{k} \left( \frac{s_j}{(1-\phi_j)^2} + \frac{l_{j+1}}{\phi_j^2} \right) y_j^2 - \sum_{j=1}^{k} h_{j,j-1}(y_j - y_{j-1})^2 \leq 0.
\]
$y^T H y$ equals to 0 if and only if $y = 0$. The fact that the Hessian matrix is negative semidefinite implies that the log-likelihood function is concave. Therefore, the unique maximum exists.

**Theorem 2.2** Let $\phi^*_0, \phi^*_1, \ldots, \phi^*_k$ denote the MLE. The estimates $\hat{\phi}_j, j = 1, \ldots, k$ constructed from any initial value $\hat{\phi}_0$ have the following properties:

1) If $\hat{\phi}_0 < \phi^*_0$, then $\hat{\phi}_j < \phi^*_j$ and $H(\hat{\phi}_k, \hat{\phi}_{k-1}) > 0$.

2) If $\hat{\phi}_0 > \phi^*_0$, then $\hat{\phi}_j > \phi^*_j$ and $H(\hat{\phi}_k, \hat{\phi}_{k-1}) < 0$.

**Proof:** Given $t_1 > 0$ and $l_1 > 0$, the first equation in (2.13) implies that $\phi_1$ is an increasing function in $\phi_0$. Therefore, $\hat{\phi}_1 < \phi^*_1$ if $\hat{\phi}_0 < \phi^*_0$. Also from the first equation in (2.13), we have that $\phi_1 - \phi_0 = t_1 \frac{\phi_0}{l_1}$ is an increasing function in $\phi_0$, which implies that $\hat{\phi}_1 - \phi_0 < \phi^*_1 - \phi^*_0$.

Assume $\hat{\phi}_j < \phi^*_j$ and $\hat{\phi}_j - \hat{\phi}_{j-1} < \phi^*_j - \phi^*_{j-1}$. From $(j + 1)$th equation in (2.13), we have that $\phi_{j+1}$ and $\phi_{j+1} - \phi_j$ are increasing functions in $\phi_j$ and $\phi_j - \hat{\phi}_{j-1}$. Thus $\hat{\phi}_{j+1} < \phi^*_{j+1}$ and $\hat{\phi}_{j+1} - \hat{\phi}_j < \phi^*_{j+1} - \phi^*_j$. By induction, we have $\hat{\phi}_j < \phi^*_j$ for all $j = 1, \ldots, k$. Moreover, $H(\phi_k, \phi_{k-1})$ is a decreasing function in $\phi_k$ and $\phi_k - \phi_{k-1}$, which implies that $H(\hat{\phi}_k, \hat{\phi}_{k-1}) > H(\phi^*_k, \phi^*_{k-1}) = 0$. Similarly, we can prove the other direction.

With Theorem 2.2, we can apply a binary search algorithm to find $\hat{\phi}^*_j, j = 0, \ldots, k$. For any measured value $x_j$, the estimated probability is computed as

$$\hat{p}^{NPMLE}_j = \hat{\phi}^*_j - \hat{\phi}^*_{j-1}, \ j = 1, \ldots, k.$$

**2.4. Resampling Methods for Ranked Set Sampling**

Due to the complexity of the sampling procedure, resampling techniques are often used for variance estimation for estimators made from ranked set samples. Because bootstrap methods are based on sampling from an estimated CDF, we now have several options for implementation of a RSS bootstrap. In this section, we outline several methods available for a RSS bootstrap with our new CDF’s.
Modarres et al. (2006) summarized three bootstrap methods for continuous RSS: bootstrap RSS by rows (BRSSR), bootstrap RSS (BRSS), and mixed row bootstrap RSS (MRBRSS) in the context of balanced continuous RSS. The idea of BRSSR is to consider each order stratum separately. For each stratum, we assign equal probability to each row and resample with replacement to generate a sample of the same size. In the context of continuous RSS, BRSSR is equivalent to estimating the EDF of each stratum separately. Let \( \hat{F}^e_{[r]} \) denote the EDF of \( r \)-th order stratum, where 
\[
\hat{F}^e_{[r]}(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_{[r],i} \leq x\}, \quad r = 1, \ldots, H.
\]
Then generate \( n \) units from \( \hat{F}^e_{[r]} \) for \( r = 1, \cdots, H \) and combine the samples generated to form a bootstrap sample. One drawback of this method is that it doesn’t work for \( n = 1 \).

The second approach BRSS provides an alternative that also works for the single cycle RSS. Instead of estimating the CDF of each stratum separately, BRSS obtains the EDF \( \hat{F} \) of the overall population first. Then bootstrap RSS is simulated from \( \hat{F} \) by using procedures in section (1.2) with the same \( H \) and \( n \) that was used in the original design.

BRSSR and BRSS, however, don’t use the partial ordering information in the ranked set sample, i.e. the units in \( r \)-th stratum have a higher probability of being less than the units in \( s \)-th stratum than being greater than the units in \( s \)-th stratum for \( r < s \). Based on this concern, Modarres et al. (2006) proposed another method, called MRBRSS. MRBRSS is different with BRSS only in generating the original samples. Instead of drawing \( H \) units from \( \hat{F}^e \) in the first step in (1.2), MRBRSS draws \( H \) units by choosing one unit independently from each \( \hat{F}^e_{[r]} \) and then ranks them to get the measured unit. Because this method generates samples from \( \hat{F}^e_{[r]} \), it doesn’t work for \( n = 1 \) either.

Although all three methods are discussed in the context of balanced RSS with continuous underlying distribution in Modarres et al. (2006), they can be easily extended to discrete RSS or unbalanced RSS. But none of the methods uses information about ties. Therefore we explore how to incorporate tie information into the bootstrap procedure.

An intuitive way to use tie information is to treat \( \hat{F}^f \) or \( \hat{F}^NPMLE \) as the parent distribution instead of \( \hat{F}^e \) in BRSS. These new methods incorporate the tie information into the
estimated CDF. In the following discussion, we call the new methods BRSSF and BRSSM, respectively.

One potential problem of BRSSM is to deal with the categories $x < x_1$ and $x > x_k$. One possible solution is that we can artificially create two possible categories $\hat{x}_0$ and $\hat{x}_{k+1}$. For example, an ad-hoc approach is to set $\hat{x}_0 = x_1$ and $\hat{x}_{k+1} = x_k$. The potential result of using this combination is that it will artificially increase the probability of sampling $x_1$ and $x_k$ and therefore the observed number of units tied to $x_1$ or $x_k$. So our suggestion is to use a second choice: that is $\hat{x}_0 = x_1 - \delta$ and $\hat{x}_{k+1} = x_k + \delta$, where $\delta$ is the smallest increment observed from $\{x_1, \ldots, x_k\}$.

Suppose the parameter of interest is $T(F)$. Then the general bootstrap procedure for estimating the standard error of the estimator $T(\hat{F})$ is summarized as follows:

1. Obtain the estimate $\hat{T}$ from the original sample $\{X_{[r],i}, l_{[r],i}, t_{[r],i}\}$.
2. Generate $B$ (eg. $B = 50$, or $100$) bootstrap ranked set samples $\{X_{[r],i}^{(b)}, l_{[r],i}^{(b)}, t_{[r],i}^{(b)}\}, b = 1, \ldots, B$ by using any bootstrap method.
3. Compute $\hat{T}^{(b)}$ from each bootstrap sample.
4. Then the standard error of $T(\hat{F})$ can be estimated as

$$\hat{SE} = \sqrt{\frac{1}{B-1} \left( \bar{\hat{T}}^{(b)} - \bar{\hat{T}} \right)^2},$$

where $\bar{\hat{T}} = \frac{1}{B} \sum_{b=1}^{B} \hat{T}^{(b)}$.

By using the estimated standard error, we can construct the 95% ($\alpha = 5\%$) confidence interval

$$(\hat{T} - z_{(1-\alpha/2)} \hat{SE}, \hat{T} + z_{(1-\alpha/2)} \hat{SE}).$$

An alternative approach to $z$-confidence interval is to use the empirical confidence interval. The resulting confidence interval is $(\hat{T}_{(\alpha/2)}, \hat{T}_{(1-\alpha/2)})$, where $\hat{T}_{(\alpha/2)}$ and $\hat{T}_{(1-\alpha/2)}$ denote the $\alpha/2$ and $1 - \alpha/2$ percentile of $\hat{T}^{(b)}, b = 1, \ldots, B$. To obtain a more accurate empirical confidence interval, a relatively large $B$ (1000 or more) is required.
The resampling procedures can be used to make inference on statistical quantities. In Section 3.4.3, we will apply these bootstrap methods to make inference on the population mean. In Section 4.3.2, we will apply these bootstrap methods to make inference on the population variance.

2.5. Simulation Study

In this section, we conduct a simulation study to compare the performance of the EDF, Frey’s estimator, and the NPMLE for some common discrete distributions. The performance of a CDF estimator is evaluated in terms of mean integrated squared error (MISE). MISE of an estimator \( \hat{F} \) is computed as the sum of mean square errors over all support points, i.e.

\[
MISE(\hat{F}) = E \sum_{x \in S} (f(x) - \hat{f}(x))^2 = \sum_{x \in S} MSE(\hat{f}(x)).
\] (2.14)

Let \( \hat{F}^{SRS} \) denote the EDF of \( F \) from a simple random sample having the same number of measured units as the ranked set sample. Then the relative efficiency of an estimator \( \hat{F} \) is defined as the ratio of MISE of \( \hat{F} \) to MISE of \( \hat{F}^{SRS} \), i.e.

\[
RE(\hat{F}) = \frac{MISE(\hat{F}^{SRS})}{MISE(\hat{F})}.
\]

In our simulation, RE is estimated by 10000 replications.

Figures 2.3, 2.4 and 2.5 exhibit the simulated relative efficiency of \( \hat{F}^e \), \( \hat{F}^{Frey} \), and \( \hat{F}^{NPMLE} \) for set sizes \( H = 3, 5 \) and cycle size \( n = 3, 5, 10 \). The distributions used in generating the sample include Discrete uniform distribution (N), Binomial distribution(10, \( p \)) and Poisson distribution (\( \lambda \)). When constructing those three estimators, we don’t use any information about the sample generation distribution.

From Figures 2.3, 2.4 and 2.5, \( \hat{F}^{NPMLE} \) always has the highest RE, which suggests \( \hat{F}^{NPMLE} \) is the most efficient estimator in terms of MISE. Moreover, \( \hat{F}^{Frey} \) is more efficient than \( \hat{F}^e \) as a result of utilizing the information about ties. The RE’s of all estimators are
greater than 1, which imply that they are more efficient than the SRS estimator for the balanced ranked set sample.

Figure 2.3. Relative efficiency of CDF estimators for discrete uniform distributions
Figure 2.4. Relative efficiency of CDF estimators for binomial distributions
Figure 2.5. Relative efficiency of CDF estimators for Poisson distributions
2.6. Conclusion

In this section, we proposed two new estimators of the CDF and pmf under discrete RSS. Frey’s Estimator assigns equal weights to all tied values within each ranked set and the weighted CDF is computed within each order stratum. It has been shown to be a more efficient estimator compared to the EDF via simulation studies. The NPMLE was proposed by incorporating information into the likelihood function. Then the optimal solution can be found via a numerical method. By simulation studies, it showed a significant improvement over the other two estimators in terms of MISE. Moreover, another advantage of the NPMLE is that it can be easily adapted to the unbalanced case.

However, there are still some limitations of the new estimators. First, Frey’s estimator doesn’t work well for unbalanced cases. Moreover, from the simulation studies, Frey’s estimator only have slight improvements over the EDF under a lot of settings. Although the NPMLE improved the RE significantly, it required the assumption that the probabilities can be only assigned to those observed values. How to relax this assumption needs further investigations.
CHAPTER 3
ON ESTIMATING THE POPULATION MEAN

3.1. Background

Estimating the population mean is one of the most studied topics in RSS research. Suppose the mean, variance, cumulative distribution function, and probability mass function of the underlying population are \( \mu, \sigma^2, F(x), \) and \( f(x) \) and the corresponding mean, variance, cumulative distribution function, and probability mass function of \( r\)-th order stratum are \( \mu_r, \sigma^2_r, F_r(x), \) and \( f_r(x). \) Let \( \{ (X_{[r]i}, l_{[r]i}, t_{[r]i}); r = 1, ..., H; i = 1, ..., n \} \) denote a balanced discrete ranked set sample, where \( l_{[r]i} \) and \( t_{[r]i} \) are the number of units which are judged to be less than the measured unit and the number of units which are judged tied to the measured unit (including the measured one itself) within the set from which \( X_{[r]i} \) is selected. Then the regular RSS mean estimator is

\[
\hat{\mu}_{\text{RSS}} = \frac{1}{H} \sum_{r=1}^{H} \hat{\mu}^{\text{RSS}}_r = \frac{1}{H} \sum_{r=1}^{H} \left( \frac{1}{n} \sum_{i=1}^{n} X_{[r]i} \right),
\]

where \( \hat{\mu}^{\text{RSS}}_r \) is the estimator of \( r\)-th stratum mean.

It is simple to verify that \( \hat{\mu}_{\text{RSS}} \) is an unbiased estimator of the population mean \( \mu. \) To compare the efficiency of \( \hat{\mu}_{\text{RSS}} \) over the SRS estimator \( \hat{\mu}_{\text{SRS}} \) from a simple random sample with the same size as the ranked set sample, we define the relative efficiency of any mean estimator \( \hat{\mu} \) as

\[
RE(\hat{\mu}) = \frac{\text{MSE}(\hat{\mu}^{\text{SRS}})}{\text{MSE}(\hat{\mu})},
\]

where \( \text{MSE}(\hat{\mu}^{\text{SRS}}) \) and \( \text{MSE}(\hat{\mu}) \) are mean square errors of SRS estimator \( \hat{\mu}^{\text{SRS}} \) and \( \hat{\mu}, \) respectively. As both \( \hat{\mu}_{\text{RSS}} \) and \( \hat{\mu}_{\text{SRS}} \) are unbiased estimators for \( \mu, \) (3.1) can be simplified.
to

$$RE(\hat{\mu}^{RSS}) = \frac{\sigma^2}{\frac{1}{H} \sum_{r=1}^{H} \sigma_r^2},$$

(3.2)

The RE of $\hat{\mu}^{RSS}$ for continuous balanced cases was proved to be bounded by 1 and $(H + 1)/2$ in Takahasi and Wakimoto (1968).

Table 3.1 exhibits the RE’s of $\hat{\mu}^{RSS}$ for some common discrete underlying distributions when $H = 2, 3, 4, 5$. The relative efficiency is still bounded by 1 and $(H + 1)/2$ for the distributions shown in the Table 3.1. Whether this upper bound holds for all discrete distributions has not been determined.

Table 3.1. Relative Efficiency of $\hat{\mu}^{RSS}$ from a balanced discrete ranked-set sample for some common discrete distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameter</th>
<th>$H = 2$</th>
<th>$H = 3$</th>
<th>$H = 4$</th>
<th>$H = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform ($N_0$)</td>
<td>3</td>
<td>1.421</td>
<td>1.801</td>
<td>2.146</td>
<td>2.466</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.455</td>
<td>1.880</td>
<td>2.287</td>
<td>2.671</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.471</td>
<td>1.920</td>
<td>2.359</td>
<td>2.780</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.483</td>
<td>1.980</td>
<td>2.463</td>
<td>2.941</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.498</td>
<td>1.995</td>
<td>2.491</td>
<td>2.985</td>
</tr>
<tr>
<td>Bin(10, p)</td>
<td>.1</td>
<td>1.387</td>
<td>1.734</td>
<td>2.049</td>
<td>2.339</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>1.444</td>
<td>1.857</td>
<td>2.246</td>
<td>2.615</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>1.450</td>
<td>1.872</td>
<td>2.271</td>
<td>2.652</td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>1.444</td>
<td>1.857</td>
<td>2.246</td>
<td>2.615</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td>1.387</td>
<td>1.734</td>
<td>2.049</td>
<td>2.339</td>
</tr>
<tr>
<td>Poisson</td>
<td>1</td>
<td>1.378</td>
<td>1.718</td>
<td>2.029</td>
<td>2.316</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.424</td>
<td>1.815</td>
<td>2.181</td>
<td>2.527</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.450</td>
<td>1.874</td>
<td>2.278</td>
<td>2.669</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.437</td>
<td>1.839</td>
<td>2.214</td>
<td>2.567</td>
</tr>
</tbody>
</table>

An interesting fact observed from Table 3.1 is that the relative efficiency of $\hat{\mu}^{RSS}$ approaches the upper bound $(H + 1)/2$ as the parameter of discrete uniform distribution $N_0 \to \infty$. Intuitively, the limiting distribution of discrete uniform distribution as $N_0 \to \infty$
is close to the continuous uniform distribution which has been shown to have the relative efficiency of \((H + 1)/2\).

### 3.2. Frey’s Estimator of the Population Mean

Note that \(\hat{\mu}^{RSS}\) doesn’t use any tie information. Frey (2012) proposed an estimator, denoted by \(\hat{\mu}^{Frey}\), which incorporates the tie information in the context of continuous JPS. The same idea can be applied in discrete RSS. This estimator only works for the balanced discrete RSS.

Suppose we have a balanced ranked set sample \(\{(X_{[r]}[i], l_{[r]}[i], t_{[r]}[i]), r = 1, \ldots, H; i = 1, \ldots, n\}\). Let \(x_1 < x_2 < \cdots < x_k\) denote \(k\) distinct measured values. Define the index set \(S_r = \{(h, i) | l_{[h]}[i] < r \leq l_{[h]}[i] + t_{[h]}[i]\}\). Then \(\hat{\mu}^{Frey}\) is

\[
\hat{\mu}^{Frey} = \frac{1}{H} \sum_{r=1}^{H} \hat{\mu}^{Frey}_{[r]}, \tag{3.3}
\]

where \(\hat{\mu}^{Frey}_{[r]}\) is the estimate of the mean of \(r\)-th stratum

\[
\hat{\mu}^{Frey}_{[r]} = \frac{\sum_{(h,i) \in S_r} X_{[h]}[i]}{\sum_{(h,i) \in S_r} t_{[h]}[i]}. \tag{3.4}
\]

Frey (2012) showed via simulation studies that \(\hat{\mu}^{Frey}\) is more efficient than the regular mean estimator \(\hat{\mu}^{RSS}\) in the context of continuous JPS.

Liu (2016) proposed a modified Horvitz-Thompson estimator \(\hat{\mu}^{HT}\) for discrete RSS. Let \(\pi_{[r]}[i]\) be the inclusion probability of \(X_{[r]}[i]\), which is

\[
\pi_{[r]}[i] = \frac{t_{[r]}[i]}{H}
\]

for balanced cases. For the unbalanced cases, \(\pi_{[r]}[i] = \frac{\sum_{(h,i) \in S_r} n_r}{N}\), where \(n_r\) is the number of units selected from the \(r\)-th ranked class. Let \(\hat{\mu}^{HT}\) and \(\hat{\mu}^{HT}_{[r]}\) be the modified Horvitz-
Thompson estimator of the population mean and $r$-th stratum mean, where

$$\hat{\mu}_{HT} = \frac{1}{H} \sum_{r=1}^{H} \hat{\mu}_{HT}^{[r]} \quad (3.5)$$

and

$$\hat{\mu}_{HT}^{[r]} = \frac{\sum_{(h,i) \in S_r \cap [h_{[k]}]} X_{[k]}}{\sum_{(h,i) \in S_r} \pi_{[h]}}. \quad (3.6)$$

They showed that their estimator was equivalent to $\hat{\mu}_{Frey}$ under balanced discrete RSS, which explained why Frey’s estimator performed so well in balanced discrete cases.

### 3.3. Nonparametric Maximum Likelihood Estimator of $\mu$

In this section, we propose a new plug-in estimator of the mean $\hat{\mu}_{NPMLE}$ by using the nonparametric maximum likelihood estimator (NPMLE) of the CDF. It is well-known that $\mu$ is a functional of the CDF $F$ for discrete distributions, i.e.

$$\mu = T(F) = \sum_{x \in S} x(F(x) - F(x^-)),$$

where $S$ is the support of the distribution. Many estimators can be constructed using this relationship. For example, $\hat{\mu}^{RSS}$ is the plug-in estimator by using $\hat{F}^e$, i.e. $\hat{\mu}^{RSS} = T(\hat{F}^e)$, and $\hat{\mu}^{Frey}$ is the plug-in estimator by using $\hat{F}^{Frey}$, i.e. $\hat{\mu}^{Frey} = T(\hat{F}^{Frey})$.

Now let us construct $\hat{\mu}_{NPMLE}$ by using $\hat{F}^{NPMLE}$. Let $x_1 < \cdots < x_k$ denote the $k$ distinct measured values in the ranked set sample. $\hat{\mu}_{NPMLE}$ is defined as

$$\hat{\mu}_{NPMLE} = T(\hat{F}^{NPMLE}) = \hat{x}_0 \hat{\phi}_0 + \sum_{j=1}^{k} x_j (\hat{\phi}_j - \hat{\phi}_{j-1}) + \hat{x}_{k+1} (1 - \hat{\phi}_k) \quad (3.7)$$

where $\hat{\phi}_0, \cdots, \hat{\phi}_k$ is the approximation of the NPMLE of the CDF from section (2.3), $\hat{x}_0$ is an estimate of values which are less than $x_1$, and $\hat{x}_{k+1}$ is an estimate of values which are greater than $x_k$. Similar to 2.4, there are many choices of $\hat{x}_0$ and $\hat{x}_{k+1}$. For example, one reasonable choice is that $\hat{x}_0 = x_1$ and $\hat{x}_{k+1} = x_k$. Another reasonable choice is that $\hat{x}_0 = x_1 - \delta$ and
\[ \hat{x}_{k+1} = x_k + \delta, \] where \( \delta \) is the smallest increment observed from \( \{x_1, \ldots, x_k\} \). In the following discussion, we will always use \( \hat{x}_0 = x_1 \) and \( \hat{x}_{k+1} = x_k \) to construct \( \hat{\mu}^{NPMLE} \).

### 3.4. Simulation Study

In this section, we examine the performance of the proposed estimators via simulation studies. There are three simulation studies in this section: Section 3.4.1 is to compare the performance of those estimators for balanced discrete RSS; Section 3.4.2 is to show the performance of \( \hat{\mu}^{NPMLE} \) for unbalanced RSS; Section 3.4.3 is to examine whether the bootstrap method proposed in Section 2.4 can be used for accurate inference on the mean. In all simulations, when computing \( \hat{\mu}^{NPMLE} \), we use that \( \hat{x}_0 = x_1 \) and \( \hat{x}_{k+1} = x_k \) in (3.7). For the bootstrap procedure from \( \hat{F}^{NPMLE} \), we increase \( \hat{\phi}_1 \) by \( \hat{\phi}_0 \) and also increase \( \hat{\phi}_k \) by \( \hat{\phi}_{k+1} \).

#### 3.4.1. Performance of Estimators Under Balanced Discrete Ranked Set Sampling

This simulation is designed to compare those estimators under balanced discrete RSS in terms of relative efficiency (\( RE \)). Let \( \hat{\mu}^{SRS} \) be the mean estimator from a simple random sample which has the same sample size as the ranked set sample. The \( RE \) of an estimator \( \hat{\mu} \) is defined as the ratio of the mean square error (MSE) of \( \hat{\mu}^{SRS} \) to that of \( \hat{\mu} \), i.e.

\[
RE(\hat{\mu}) = \frac{\text{MSE}(\hat{\mu}^{SRS})}{\text{MSE}(\hat{\mu})}. \tag{3.8}
\]

In our simulation studies, \( \text{MSE}(\hat{\mu}) \) is estimated based on 10000 simulated samples for each combinations of parameters.

We set the set size (\( H \)) as 3 and 5. The cycle size (\( n \)) is set to be 3, 5, and 10. For the underlying distribution, we choose three common discrete distributions, including Discrete Uniform distribution (\( \text{Uniform}(N_0) \)), Poisson distribution (\( Poi(\lambda) \)), and binomial distribution (\( Bin(n = 10, p) \)). Then we examined the estimators for a range of parameters settings.
for each distribution. The results for the three distributions are shown in Figures 3.1, 3.2, and 3.3, respectively.

Figure 3.1. Relative efficiency of mean estimators for discrete uniform distributions ($N_0$).
Figure 3.2. Relative efficiency for binomial distributions \((n = 10, p)\).
Figure 3.3. Relative efficiency of mean estimators for Poisson distributions ($\lambda$).
From Figures 3.1, 3.2, and 3.3, we see that the RE’s of all RSS estimators are uniformly greater than 1 for all underlying distributions, which implies that those mean estimators are more efficient than SRS estimator with the same number of measured units. Also, the relative efficiency of $\hat{\mu}_{RSS}$ is still bounded by $\frac{H+1}{2}$, which is consistent with the conclusion under continuous RSS. But for $\hat{\mu}_{Frey}$ and $\hat{\mu}_{NPMLE}$, the maximum relative efficiency may exceed $\frac{H+1}{2}$ as a result of using the information about ties. Moreover, $\hat{\mu}_{Frey}$ outperforms $\hat{\mu}_{RSS}$, which is consistent with the conclusion for continuous cases in Frey (2012). $\hat{\mu}_{NPMLE}$ outperforms $\hat{\mu}_{RSS}$ and $\hat{\mu}_{Frey}$ in every case. In other words, the proposed estimator is the most efficient estimator among those three estimators for the distributions considered.

3.4.2. Performance of Estimators Under Unbalanced Discrete Ranked Set Sampling

The RE of a mean estimator with an unbalanced ranked set sample is affected by many factors, including the sampling proportion from each stratum, underlying distribution, etc. To examine the effect of unbalance, we chose a simulation setting of $H = 2$ and $N = 30$. The sample size from the first ranked stratum varies from 1 to 29. Then to study the effect of varying distributions, we simulated these unbalanced ranked set samples from six distributions: discrete uniform($N_0 = 5, 10$), Poisson($\lambda = 1, 3$) and binomial($n = 10, p = .2, .5$). Each setting of distribution and sample allocation was replicated 10000 times, and the four estimators of mean $\hat{\mu}_{RSS}, \hat{\mu}_{Frey}, \hat{\mu}_{HT}$, and $\hat{\mu}_{NPMLE}$ computed from each. Then the empirical bias, variance, MSE, and RE of each estimator were computed from the replicates.

Figure 3.4 shows the RE as a function of the sampling proportion from 1-st ranking class for each distribution. Table 3.2 showed the simulated bias, variance and MSE of each mean estimator for $binomial(10, 0.2)$. 

Figure 3.4. Relative efficiency of mean estimators versus sampling proportion from 1st order stratum for discrete uniform ($N_0 = 5, 10$), Poisson ($\lambda = 1, 3$) and binomial ($n = 10, p = .2, .5$) for $H = 2$. 
From Figure 3.4, \( \hat{\mu}^{NPMLE} \) is the most efficient estimator compared to \( \hat{\mu}^{RSS} \), \( \hat{\mu}^{Frey} \), and \( \hat{\mu}^{HT} \) for all distributions considered. It always outperforms the SRS estimator as the RE is always greater than 1. Also, the maximum RE of \( \hat{\mu}^{NPMLE} \) is not always achieved when sampling proportion from 1st stratum is 0.5, which suggests that \( \hat{\mu}^{NPMLE} \) for unbalanced RSS may have a higher RE than \( \hat{\mu}^{NPMLE} \) for balanced RSS as would be expected, based on the theory of Neyman allocation.

But \( \hat{\mu}^{Frey} \), \( \hat{\mu}^{HT} \), and \( \hat{\mu}^{RSS} \) may not always outperform the SRS estimator, especially when the proportion is close to 0 or 1. Moreover, \( \hat{\mu}^{Frey} \) and \( \hat{\mu}^{HT} \) are more efficient than \( \hat{\mu}^{RSS} \). \( \hat{\mu}^{Frey} \) and \( \hat{\mu}^{HT} \) have similar MSE’s and they have exactly the same RE’s when sampling proportion \( p = .5 \) because they are equivalent when \( p = .5 \). Although MSE’s of \( \hat{\mu}^{Frey} \) and \( \hat{\mu}^{HT} \) are similar for all the sampling proportions, the source of the components of the MSE are different. As shown in Table 3.2, \( \hat{\mu}^{Frey} \) may have a serious bias when sampling proportion is close to 0 or 1, whereas the bias of \( \hat{\mu}^{HT} \) is always small. In other words, \( \hat{\mu}^{HT} \) has a relatively high variance compared to \( \hat{\mu}^{Frey} \) when the sampling proportion is close to 0 or 1 because \( \hat{\mu}^{HT} \) correctly differentially weights the observations from the two strata based on the probability that the units are observed, while \( \hat{\mu}^{Frey} \) does not.
Table 3.2. Simulated bias, variance and MSE of mean estimators for Binomial (10, 0.2) based on 10000 replicates ($N = 30$)

<table>
<thead>
<tr>
<th>Proportion</th>
<th>$\hat{\mu}^{\text{Frey}}$ Bias</th>
<th>Variance</th>
<th>MSE</th>
<th>$\hat{\mu}^{\text{HT}}$ Bias</th>
<th>Variance</th>
<th>MSE</th>
<th>$\hat{\mu}^{\text{NPMLE}}$ Bias</th>
<th>Variance</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.164</td>
<td>0.044</td>
<td>0.071</td>
<td>0.001</td>
<td>0.069</td>
<td>0.020</td>
<td>0.034</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.103</td>
<td>0.038</td>
<td>0.048</td>
<td>0.002</td>
<td>0.043</td>
<td>0.008</td>
<td>0.033</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.062</td>
<td>0.033</td>
<td>0.038</td>
<td>0.001</td>
<td>0.035</td>
<td>0.003</td>
<td>0.032</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.026</td>
<td>0.032</td>
<td>0.033</td>
<td>-0.004</td>
<td>0.032</td>
<td>-0.004</td>
<td>0.032</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.000</td>
<td>0.033</td>
<td>0.033</td>
<td>0.000</td>
<td>0.033</td>
<td>-0.001</td>
<td>0.033</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>-0.032</td>
<td>0.035</td>
<td>0.036</td>
<td>0.000</td>
<td>0.035</td>
<td>-0.002</td>
<td>0.033</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>-0.073</td>
<td>0.038</td>
<td>0.043</td>
<td>-0.002</td>
<td>0.041</td>
<td>-0.007</td>
<td>0.035</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>-0.130</td>
<td>0.043</td>
<td>0.060</td>
<td>-0.004</td>
<td>0.053</td>
<td>-0.014</td>
<td>0.036</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.226</td>
<td>0.053</td>
<td>0.104</td>
<td>-0.010</td>
<td>0.097</td>
<td>-0.033</td>
<td>0.041</td>
<td>0.042</td>
<td></td>
</tr>
</tbody>
</table>

3.4.3. Inference on Population Mean

We conducted a simulation study designed to examine whether the bootstrap method proposed in Section 2.4 can be used for accurate inference on the mean. Specifically, we examine the performance of the bootstrap based estimator of standard error and the coverage probability of the confidence interval procedures produced from that estimator, as outlined in Section 2.4.

The simulation was conducted as a factorial experiment with three factors: distribution, set size, and cycle size. Samples were simulated from six underlying distributions: discrete uniform(5), discrete uniform(10), binomial(10, 0.2), binomial(10,.5), Poisson(1), and Poisson(5). From each distribution, we generated $R = 2000$ balanced ranked set samples. We examined two settings of set size and three settings of number of cycles: $H = 3$ and 5, and $n = 3, 5,$ and 10.

From each sample, we computed $\hat{\mu}^{\text{NPMLE}}$ and its standard error using the BRSSM bootstrap procedure. That is, we computed $\hat{F}^{\text{NPMLE}}$, from which we simulated $B = 100$
bootstrap ranked set samples with the same ranked set sample design as its parent sample. From each sample, we computed \( \hat{\mu}_b^{(b)NPMLE} \) for \( b = 1, \cdots, 100 \). Then the standard error of \( \hat{\mu}^{NPMLE} \) was estimated by

\[
\hat{SE}(\hat{\mu}^{NPMLE}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\mu}_b^{(b)NPMLE} - \bar{\hat{\mu}}^{NPMLE})^2},
\]

where

\[
\bar{\hat{\mu}}^{NPMLE} = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b^{(b)NPMLE}.
\]

For each of the samples, a 95% confidence interval for \( \mu \) was also constructed as follows:

\[
(\hat{\mu}^{NPMLE} - z_{0.975} \times \hat{SE}, \hat{\mu}^{NPMLE} + z_{0.975} \times \hat{SE}).
\]

(3.10)

To evaluate the results of the simulation, we computed the empirical bias of the standard error estimator (3.9) and the coverage probability of the confidence intervals (3.10). To examine the bias of \( \hat{SE}(\hat{\mu}^{NPMLE}) \), we first computed the average of \( \hat{SE}(\hat{\mu}^{NPMLE}) \) for the \( R = 2000 \) replicates for each simulating setting which we denote by \( \bar{\hat{SE}}(\hat{\mu}^{NPMLE}) \). Then we compared this average to the simulated standard error of \( \hat{\mu}^{NPMLE} \), denoted by \( SE^* \). \( SE^* \) is simply the sample standard deviation of the \( R = 2000 \) replicates of \( \hat{\mu}^{NPMLE} \). To investigate the performance of the confidence intervals, we computed the proportion of intervals which cover the mean of the relevant parent distribution.

The results of this evaluation are shown in Table 3.3. For each of the scenarios considered, we display \( \bar{\hat{SE}}, SE^* \), and the confidence interval coverage proportion. These are shown in columns \( \bar{\hat{SE}}, SE^* \), and Coverage of the table.

We first note that the standard errors and their estimates do behave as we would expect based on the theory of ranked set sampling. They decrease with increasing sample size, and they show the effect of greater precision as \( H \) increases. For example, we observe that \( \bar{\hat{SE}} \) and \( SE^* \) are both smaller for the cases of \( H = 5, n = 3 \) than for \( H = 3, n = 5 \), as theory would predict.
To examine the performance of the bootstrap procedure for inference, notice that the standard error estimator appears to have small bias. That is, $\hat{SE}$ and $SE^*$ are usually quite close to each other. The most noticeable bias occurs for the binomial and Poisson distributions, where the biases are all negative, and most noticeable for small sample sizes. The reason for this could be that a support point at the endpoint of the support range is less likely to be observed in a small sample, resulting in an underestimate of the standard error. The binomial and Poisson distributions do have relatively smaller probability masses near the edges of their range.

Turning to the confidence intervals, we see that for all scenarios, the coverage probabilities are close to their nominal level of 95%. The slight undercoverage of confidence intervals for the Binomial and Poisson distributions, especially for small sample size, is presumably due to the small negative bias of the estimators of the standard errors.
Table 3.3. Coverage probability, the average of estimates of the standard error of $\hat{\mu}^{NPMLE}$, and the simulated standard error of $\hat{\mu}^{NPMLE}$ based on 2000 replicates

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$n$</th>
<th>$\bar{SE}$</th>
<th>$SE^*$</th>
<th>Coverage</th>
<th>$\bar{SE}$</th>
<th>$SE^*$</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform (5)</td>
<td>3</td>
<td>0.064</td>
<td>0.063</td>
<td>0.928</td>
<td>0.038</td>
<td>0.038</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.047</td>
<td>0.048</td>
<td>0.933</td>
<td>0.029</td>
<td>0.029</td>
<td>0.944</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.033</td>
<td>0.033</td>
<td>0.957</td>
<td>0.020</td>
<td>0.021</td>
<td>0.943</td>
</tr>
<tr>
<td>uniform (10)</td>
<td>3</td>
<td>0.072</td>
<td>0.0657</td>
<td>0.933</td>
<td>0.044</td>
<td>0.040</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.052</td>
<td>0.051</td>
<td>0.928</td>
<td>0.032</td>
<td>0.032</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.035</td>
<td>0.035</td>
<td>0.954</td>
<td>0.022</td>
<td>0.023</td>
<td>0.942</td>
</tr>
<tr>
<td>Binomial(10, 0.2)</td>
<td>3</td>
<td>0.269</td>
<td>0.289</td>
<td>0.913</td>
<td>0.174</td>
<td>0.177</td>
<td>0.928</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.212</td>
<td>0.223</td>
<td>0.924</td>
<td>0.134</td>
<td>0.138</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.150</td>
<td>0.153</td>
<td>0.943</td>
<td>0.095</td>
<td>0.096</td>
<td>0.946</td>
</tr>
<tr>
<td>Binomial(10, 0.5)</td>
<td>3</td>
<td>0.340</td>
<td>0.360</td>
<td>0.924</td>
<td>0.222</td>
<td>0.221</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.268</td>
<td>0.281</td>
<td>0.916</td>
<td>0.170</td>
<td>0.175</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.191</td>
<td>0.195</td>
<td>0.948</td>
<td>0.121</td>
<td>0.124</td>
<td>0.947</td>
</tr>
<tr>
<td>Poisson(1)</td>
<td>3</td>
<td>0.207</td>
<td>0.226</td>
<td>0.893</td>
<td>0.133</td>
<td>0.136</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.163</td>
<td>0.174</td>
<td>0.921</td>
<td>0.104</td>
<td>0.108</td>
<td>0.928</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.117</td>
<td>0.119</td>
<td>0.936</td>
<td>0.074</td>
<td>0.076</td>
<td>0.931</td>
</tr>
<tr>
<td>Poisson(5)</td>
<td>3</td>
<td>0.481</td>
<td>0.522</td>
<td>0.910</td>
<td>0.317</td>
<td>0.324</td>
<td>0.929</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.387</td>
<td>0.411</td>
<td>0.914</td>
<td>0.247</td>
<td>0.255</td>
<td>0.926</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.274</td>
<td>0.280</td>
<td>0.947</td>
<td>0.177</td>
<td>0.180</td>
<td>0.947</td>
</tr>
</tbody>
</table>

3.5. Conclusion

In this section, we proposed a new plug-in estimator of the mean based on the NPMLE of the CDF. The proposed estimator has been shown to be more efficient than those existing estimators via simulation studies. Moreover, the new proposed estimator has an additional benefit that it can be extended to unbalanced RSS because the likelihood function of the ranked set sample doesn’t require the sample is to be from a balanced design. A simulation study for unbalanced RSS when $H = 2$ suggests that the new proposed estimator is the only
estimator which may consistently outperform the SRS estimator. We also showed that the bootstrap procedure can be used to make inference on $\mu$ in balanced discrete RSS.
CHAPTER 4
ON ESTIMATING THE POPULATION VARIANCE

4.1. Background

It is common that the investigators are also interested in estimating the population variance. In this section, we focus on estimating the population variance \( \sigma^2 \) under discrete RSS. Suppose the mean, variance, cumulative distribution function, and probability mass function of the underlying population are \( \mu, \sigma^2, F(x), \) and \( f(x) \) and the corresponding mean, variance, cumulative distribution function, and probability mass function of \( r \)-th order stratum are \( \mu_{[r]}, \sigma_{[r]}^2, F_{[r]}(x), \) and \( f_{[r]}(x) \). Let \( \{(X_{[r]i}, l_{[r]i}, t_{[r]i}); r = 1, \ldots, H; i = 1, \ldots, n\} \) denote a balanced discrete ranked set sample with set size \( H \) and cycle size \( n \), where \( l_{[r]i} \) and \( t_{[r]i} \) are the number of units which are judged to be less than the measured unit and the number of units which are judged tied to the measured unit (including the measured one itself) within the set from which \( X_{[r]i} \) is selected.

Here, we first review some existing estimators which were proposed for estimating the variance in the context of continuous RSS. The first estimator was proposed by Stokes (1980). For a balanced ranked set sample \( \{X_{[r]i}\} \), Stokes’s estimator is

\[
\hat{\sigma}_{stokes}^2 = \frac{1}{nH - 1} \sum_{i=1}^{n} \sum_{r=1}^{H} (X_{[r]i} - \bar{\mu}_{RSS})^2,
\]

where \( \bar{\mu}_{RSS} = \frac{1}{nH} \sum_r \sum_i X_{[r]i} \). Although this estimator is not unbiased, it is asymptotically unbiased (which can be shown directly from (4.1)) and asymptotically more efficient than that of the variance estimator from a simple random sample with the same number of
measured units. The bias and variance of $\sigma^2_{\text{stokes}}$ were given by

$$\text{bias}(\hat{\sigma}^2_{\text{stokes}}) = \frac{\sum_r (\mu_r - \mu)^2}{H(nH - 1)}$$

$$\text{Var}(\hat{\sigma}^2_{\text{stokes}}) = \frac{1}{nH^2} \sum_r \mu^4r + \frac{4n}{(nH - 1)^2} \sum_r \tau^2_r \sigma^2_r + \frac{4n}{nH - 1} \sum_r \tau_r \mu^3_r$$

$$+ \frac{4}{H^2(nH - 1)^2} \sum_r \sum_{r<s} \sigma^2_r \sigma^2_s - \frac{2(n - 1) - (nH - 1)^2}{nH^2(nH - 1)^2} \sum_r \sigma^4_r,$$

where $\mu_{j[r]} = E(X_{[r]} - \mu_{[r]})^j$ and $\tau_{[r]} = \mu_{[r]} - \mu$.

MacEachern et al. (2002) proposed an unbiased estimator of the population variance, which is

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{\sum_{r\neq s} \sum_i \sum_j (X_{[r]i} - X_{[s]j})^2}{2n^2H^2} + \frac{\sum_{r=s} \sum_i \sum_j (X_{[r]i} - X_{[s]j})^2}{2n(n - 1)H^2}.$$

and the variance of $\hat{\sigma}^2_{\text{unbiased}}$ is given by

$$\text{Var}(\hat{\sigma}^2_{\text{unbiased}}) = \frac{1}{nH^2} \sum_r \mu^4r + \frac{4n}{nH^2} \sum_r \tau^2_r \sigma^2_r + \frac{4n}{nH - 1} \sum_r \tau_r \mu^3_r$$

$$+ \frac{4}{n^2H^4} \sum_r \sum_{r<s} \sigma^2_r \sigma^2_s - \frac{H^2(n - 1) - 2}{nH^4(n - 1)} \sum_r \sigma^4_r.$$

Perron and Sinha (2004) independently proposed the same estimator and also proved that it is a minimum variance unbiased quadratic estimator. $\hat{\sigma}^2_{\text{stokes}}$ and $\hat{\sigma}^2_{\text{unbiased}}$ can be applied under discrete RSS by ignoring the tie information.

### 4.2. New Estimators of Population Variance

In this section, we propose two new estimators of $\sigma^2$ under discrete RSS by using the fact that

$$\sigma^2 = \sum_{x \in S} (x - \mu)^2 f(x),$$

(4.2)
where $S$ is the support of the distribution, $\mu$ is the population mean, and $f(x)$ is probability mass function. Given any estimate of $F$, say $\hat{F}$, the plug-in estimator $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \sum_{x \in S_0} (x - \mu(\hat{F}))^2(\hat{F}(x) - \hat{F}(x-)),$$

where $S_0$ is the set of all measured values and $\mu(\hat{F})$ is the plug-in estimate of $\mu$ using $\hat{F}$.

In the context of continuous RSS, it is trivial to show that with the empirical estimators $\hat{\mu}^{RSS}$ and $\hat{F}^e$, we obtain

$$\hat{\sigma}^2_e = \sum_{x \in S_0} (x - \hat{\mu}^{RSS})^2(\hat{F}^e(x) - \hat{F}^e(x-)) = \frac{1}{nH} \sum_{i=1}^n \sum_{r=1}^H (X_{[r]i} - \hat{\mu}^{RSS})^2.$$

$\hat{\sigma}^2_{stokes}$ is the corrected version of $\hat{\sigma}^2_e$ with a correction factor $\frac{nH}{nH-1}$.

By using $\hat{F}^{Frey}$ and $\hat{F}^{NPMLE}$ from Chapter 2, we can obtain the corresponding variance estimator $\hat{\sigma}^2_{Frey}$ and $\hat{\sigma}^2_{NPMLE}$. As $\hat{F}^{Frey}$ and $\hat{F}^{NPMLE}$ have been shown to be more efficient than $\hat{F}^e$, they are expected to produce more efficient variance estimators.

### 4.3. Simulation Study

In this section, we describe simulation studies we conducted to investigate the relative efficiency of $\hat{\sigma}^2_{Frey}$ and $\hat{\sigma}^2_{NPMLE}$ compared to that of $\hat{\sigma}^2_{stokes}$ and $\hat{\sigma}^2_{unbiased}$. We also report findings from an examination of the performance of the bootstrap method for estimating the standard error of $\hat{\sigma}^2_{NPMLE}$ and constructing confidence interval based on $\hat{\sigma}^2_{NPMLE}$.

#### 4.3.1. Comparison of Variance Estimators

Let $\hat{\sigma}^2_{SRS}$ be the variance estimator from a simple random sample which has the same sample size as the ranked set sample. The relative efficiency of an estimator $\hat{\sigma}^2$ is defined as the ratio of the mean squared errors between the SRS estimator $\hat{\sigma}^2_{SRS}$ and $\hat{\sigma}^2$ with the same
number of measured units, i.e.

\[ RE(\hat{\sigma}^2) = \frac{MSE(\hat{\sigma}^2_{SRS})}{MSE(\hat{\sigma}^2)}. \]

In the simulation studies, \( MSE(\hat{\sigma}^2) \) and \( MSE(\hat{\sigma}^2_{SRS}) \) are estimated using 10000 replications. The sample units are generated from three common discrete distributions: discrete uniform distribution \((N_0)\), binomial distribution \((Bin(10, p))\) and Poisson distribution \((\lambda)\). For each, we simulated samples from a range of parameter values. We set the set size \((H)\) to be 3 and 5, and the number of units from each stratum to be 3, 5 and 10.

The results of the simulations are shown in Figures 4.1-4.3. Each figure shows the RE’s for two plug-in estimators of \(\hat{\sigma}^2\) \((\hat{\sigma}^2_{Frey} \text{ and } \sigma^2_{NPMLE})\) and the two previous RSS variance estimators that don’t use the tie information \((\sigma^2_{stokes} \text{ and } \sigma^2_{unbiased})\) for one distribution (discrete uniform, binomial, Poisson).
Figure 4.1. Relative efficiency of population variance estimators for discrete uniform distributions ($N_0$).
Figure 4.2. Relative efficiency of population variance estimators for binomial distributions ($n = 10$, $p$).
Figure 4.3. Relative efficiency for Poisson distributions (\(\lambda\)).
From Figure 4.1, 4.2, 4.3, it is clear that $\hat{\sigma}^2_{NPMLE}$ is consistently more efficient than the other estimators in terms of MSE. Moreover, $\hat{\sigma}^2_{Frey}$ is more efficient than $\hat{\sigma}^2_{unbiased}$ and $\hat{\sigma}^2_{stokes}$. In other words, tie information improves the estimation of the population variance and $\hat{\sigma}^2_{NPMLE}$ uses tie information more efficiently than $\hat{\sigma}^2_{Frey}$. Not surprisingly, the relative efficiency of variance estimators is higher for $H = 5$ compared to $H = 3$ when the size of total measured units is 15.

4.3.2. Inference on Population Variance

As shown by the simulation in Section 3.4.3, the bootstrap method can be used for estimating the variation and providing statistical inference on the mean. It is also of interest to see whether the same method can be used for statistical inference on the population variance.

In this simulation study, we used the same settings as presented in Section 3.4.3. The simulation study was conducted as a factorial experiment with three factors: distribution, set size, and cycle size. Samples were simulated from six underlying distributions: discrete uniform(5), discrete uniform(10), binomial(10, 0.2), binomial(10, 0.5), Poisson(1), and Poisson(5). From each distribution, we generated $R = 2000$ balanced ranked set samples. We examined two settings of set size and three settings of number of cycles: $H = 3$ and 5, and $n = 3, 5, \text{ and } 10$.

From each sample, we computed $\hat{\sigma}^2_{NPMLE}$ and its standard error using the BRSSM bootstrap procedure. That is, we computed $\hat{F}^{NPMLE}$, from which we simulated $B = 100$ bootstrap ranked set samples with the same ranked set sample design as its parent sample. From each sample, we computed $\hat{\sigma}^{2(b)}_{NPMLE}$ for $b = 1, \ldots, 100$. Then the standard error of $\hat{\sigma}^2_{NPMLE}$ was estimated by

$$\hat{SE}(\hat{\sigma}^2_{NPMLE}) = \frac{1}{B - 1} \sqrt{\sum_{b=1}^{B} \left( \hat{\sigma}^{2(b)}_{NPMLE} - \bar{\sigma}^2_{NPMLE} \right)^2},$$

(4.3)
where

\[ \hat{\sigma}^2_{NPMLE} = \frac{1}{B} \sum_{b=1}^{B} \hat{\sigma}^2_{NPMLE}(b). \]

To perform inference on the variance, we need to develop a procedure for producing a bootstrap confidence interval. We found that the normality-based confidence interval worked well for the mean. In order for that approach to work here, it would be necessary for the sampling distribution of \( \hat{\sigma}^2_{NPMLE} \) to be approximately normal. We examined this via simulation for several cases, and found that the sampling distribution may be slightly right-skewed. An example of such a sampling distribution of \( \hat{\sigma}^2_{NPMLE} \) is shown in Figure 4.4. This sampling distribution is for the variance estimates of a sample of size \( H = 5, n = 5 \) from a binomial(10, 0.5) distribution. A logarithmic transformation is often useful for data which have right skewness like this. For comparison purposes, we produced confidence intervals both without log-transformation and with log-transformation for the NPMLE variance estimate.

![Histogram of \( \hat{\sigma}^2_{NPMLE} \) based on 2000 replicates, \( H=5, n=5 \)]

Figure 4.4. Histogram of \( \hat{\sigma}^2_{NPMLE} \) when ranked set samples are generated from binomial(10, 0.5). The set size is 5 and the cycle size is 5.

For each of the samples, a 95\% \( z \)-confidence interval for \( \sigma^2 \) without taking the log-transformation was constructed as follows:
\[ (\hat{\sigma}^2_{NPMLE} - z_{0.975} \ast \hat{SE}, \hat{\sigma}^2_{NPMLE} + z_{0.975} \ast \hat{SE}). \]

(4.4)

With the log-transformation, a z-interval was computed as follows:

\[ (e^{\text{log}(\hat{\sigma}^2_{NPMLE}) - z_{0.975} \ast \hat{SE}(\text{log}(\hat{\sigma}^2_{NPMLE}))}, e^{\text{log}(\hat{\sigma}^2_{NPMLE}) + z_{0.975} \ast \hat{SE}(\text{log}(\hat{\sigma}^2_{NPMLE}))}). \]

(4.5)

where \( \hat{SE}(\text{log}(\hat{\sigma}^2_{NPMLE})) \) is the bootstrap estimate of standard error of \( \text{log}(\hat{\sigma}^2_{NPMLE}) \).

To evaluate the simulation results, we computed the empirical bias of the standard error estimator (4.3) and the coverage probability of the confidence interval procedure. To examine the bias of \( \hat{\sigma}^2_{NPMLE} \), we computed the average of \( \hat{SE}(\hat{\sigma}^2_{NPMLE}) \) for the \( R = 2000 \) replicates for each simulating setting, which we denote by \( \bar{\hat{SE}} \). Then we compared this average to the simulated standard error of \( \hat{\sigma}^2_{NPMLE} \), denoted by \( SE^* \). \( SE^* \) is simply the sample standard deviation of the \( R = 2000 \) replicates of \( \hat{\sigma}^2_{NPMLE} \). To investigate the performance of the confidence interval, we computed the proportion of intervals which cover the variance of the relevant parent distribution.

The results of this evaluation are shown in Table 4.1. For each of the scenarios considered, we display \( \bar{\hat{SE}}, SE^* \), the confidence interval coverage proportion without log-transformation, and the confidence interval coverage proportion with log-transformation. These are shown in columns \( \bar{\hat{SE}}, SE^*, \text{Cover}, \) and \( \text{Cover}_l \) of the table.

We first note that the standard errors and their estimates behave as we would expect based on the theory of ranked set sampling in which the standard errors decrease with the increase in sample size. We can also observe that, like the mean estimator, estimation is more precise for larger \( H \) and for equal total sample size. For example, we observe that \( \bar{\hat{SE}} \) and \( SE^* \) are both smaller for the cases of \( H = 5, n = 3 \) than for \( H = 3, n = 5 \), as we expected.

Table 4.1 shows that the bias in the bootstrap standard error estimate is larger for the variance than we saw for the mean estimator. For the binomial and Poisson distributions, the bias of the bootstrap estimator of standard error is negative and relatively larger than
for the mean. For the uniform distribution, the estimator of standard error is biased upward and small.

The confidence interval coverage rate is further than the nominal value than those of the mean even after taking the log-transformation. For most cases, the coverage rates of intervals after taking the log-transformation are higher than that without taking the log-transformation. The improvement is larger for the small sample size than for the large sample size. For the uniform distribution, the bias of coverage probabilities are small for both procedures. But for the binomial and Poisson distributions, the coverage probabilities of both procedures are always lower than the nominal level 0.95, presumably because the standard error estimator underestimates the standard error of $\hat{\sigma}^2_{NPMLE}$. But we also observe that the coverage probability for both procedures increases as the cycle size increases, although it never exceeds about 93% coverage even for samples of size 50 ($H = 5, n = 10$).

We conducted a more extensive simulation to better determine the required sample size for obtaining adequate coverage probability of the confidence interval. For this simulation, we kept $H = 5$, but examined the coverage probability of the confidence interval procedures as $n$ increased from 3 to 16 for binomial and Poisson distributions. The results for the binomial(10, 0.5) distribution are shown in Figure 4.5, which shows confidence interval coverage rates plotted as a function of $n$. The figure shows that even for a sample of $5 \times 16 = 80$, the coverage rate is still only about 93%.

Another noticeable feature of this plot is its oscillating slope, which is more noticeable for the intervals built without taking the log-transformation. That is, the coverage rate does not increase smoothly toward the nominal coverage probability. Both these characteristics were present for all four of the distributions examined. We did not know whether this behavior was a unique feature of the non-parametric MLE plug-in estimator for the ranked set sample, or would occur just due to the discreteness of its sampling distribution. It is well known that normality-based approximate confidence intervals for binomial proportions show this oscillating behavior (Brown et al. (2001)). Thus it perhaps should not be surprising to find this to be true also for estimation of variance from a discrete distribution.
To examine whether this behavior also occurred for a bootstrap confidence interval for variance based on the sample variance from a simple random sample, we repeated the simulation described above for that case. The results presented in Figure 4.5 are very similar to the case of ranked set samples. The oscillating behavior of the coverage rate as the sample size increases is present. Another interesting fact from Figure 4.5 is that for the small sample size, the improvement on coverage rates by taking log-transformation is larger for intervals from simple random samples than that for intervals from ranked set samples. This is presumably because the ranked set samples have a higher probability of observing more values than simple random samples, which results in a less skewed sampling distribution of the variance estimate.

![Figure 4.5](image)

Figure 4.5. Coverage probability of confidence intervals produced by the bootstrap procedure when ranked set samples or simple random samples generated for binomial(10,0.5) when $H = 5$. For simple random sample, the $x$-axis denotes equivalent cycle size; that is the sample size divided by $H$.

In summary, we have found that the bootstrap approach for estimating the standard error of $\hat{\sigma}_{NPMLE}^2$ does not work as well as it does for the mean, especially for the binomial and Poisson distributions. Part of the problem may be the fact that the size of the support
set of the random variable is underestimated when probability masses become very small at
the extreme values of the support. But that does not completely explain the undercoverage,
because it persists even for large sample sizes, for which the entire support is likely to be
represented in the sample. The coverage rates of intervals can be improved by taking log-
transformation on variance estimates. But even after taking log-transformation, the resulting
confidence intervals still undercover the true variance. We also believe that the coverage is
unlikely to improve substantially with a percentile bootstrap, because the problem does not
appear to be due to the shape of the sampling distribution of \( \hat{\sigma}_{NPMLE}^2 \).

Table 4.1. Coverage probability, the average of standard error estimates of \( \hat{\sigma}_{NPMLE}^2 \), and
the simulated standard error of \( \hat{\sigma}_{NPMLE}^2 \) based on 2000 replicates

<table>
<thead>
<tr>
<th>Distribution</th>
<th>H = 3</th>
<th>H = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n )</td>
<td>( \tilde{SE} )</td>
</tr>
<tr>
<td>uniform (5)</td>
<td>3</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.009</td>
</tr>
<tr>
<td>uniform (10)</td>
<td>3</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.011</td>
</tr>
<tr>
<td>Bin(10, 0.2)</td>
<td>3</td>
<td>0.484</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.412</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.306</td>
</tr>
<tr>
<td>Bin(10, 0.5)</td>
<td>3</td>
<td>0.751</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.627</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.463</td>
</tr>
<tr>
<td>Poisson(1)</td>
<td>3</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.279</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.218</td>
</tr>
<tr>
<td>Poisson(5)</td>
<td>3</td>
<td>1.531</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.343</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.016</td>
</tr>
</tbody>
</table>

52
4.4. Conclusion

In this chapter, we proposed two new estimators of population variance based on $\hat{F}_{\text{Frey}}$ and $\hat{F}_{\text{NPMLE}}$. The new estimators have been shown to be more efficient than its competitors $\hat{\sigma}^2_{\text{stokes}}$ and $\hat{\sigma}^2_{\text{unbiased}}$ in terms of MSE as a result of incorporating tie information.

The new variance estimators, however, share the same drawbacks of the CDF estimators. First, $\hat{\sigma}^2_{\text{Frey}}$ doesn’t work for the unbalanced RSS. Although $\hat{\sigma}^2_{\text{NPMLE}}$ can be easily extended to unbalanced RSS, it can’t incorporate any prior information of possible population values, which limits the use of this estimator.

Moreover, we also showed that the bootstrap procedure can be used to estimate the standard error of $\hat{\sigma}^2_{\text{NPMLE}}$. However, as shown in the simulation, the estimates of the standard error produced by the bootstrap procedure may underestimate the true standard error for binomial distribution and Poisson distribution. To achieve a high accuracy, we need a large sample size for those distributions.
5.1. Literature Review

In Chapter 2, we introduced three methods of estimating the CDF of the overall population. The empirical estimator and Frey’s estimator construct the estimators based on the estimated CDF of each judgment class. Most methods in the literature, actually, adopted the same idea. Therefore, a natural idea to improve the estimation efficiency is to improve the estimation efficiency for each judgment class. A common approach is to impose a constraint among judgment class distributions. For example, Ozturk (2007) introduced an estimator of \( r \)-th order stratum CDF under a stochastic ordering restriction for continuous RSS. Their proposed estimator is the minimizer of a version of the Cramér-von Mises distance function. By simulation studies, they showed that the proposed estimator for each order stratum is more efficient than the empirical estimator as measured by mean integrated squared error (MISE). This method can be applied in discrete RSS. More details will be introduced in section 5.3. Based on the idea of Ozturk (2007), Wang et al. (2012) proposed two methods to estimate the population CDF from JPS samples with empty strata. Moreover, Wang et al. (2008) proposed an estimator for the population mean by imposing a stochastic ordering on the means of the strata.

In this chapter, we consider a new estimation method by imposing a stronger constraint which is called uniform stochastic ordering on the CDF’s of different judgment classes. Dykstra et al. (1991) developed statistical inference for uniform stochastic ordering in several populations. The problem considered here is a simpler version of that in Dykstra et al. (1991).
In Section 5.2, we briefly review three different kinds of stochastic ordering. The estimator proposed by Ozturk (2007) is introduced in Section 5.3. In Section 5.4, we propose a new estimator of the CDF of each stratum under uniform stochastic ordering. In Section 5.5, we conduct a simulation study to evaluate the performance of our new estimator compared to the existing estimators.

5.2. Three Types of Stochastic Ordering

A stochastic ordering is a way to compare probability distributions. We focus on three kinds of stochastic ordering between two distributions: ordinary stochastic ordering, uniformly stochastic ordering, and likelihood ratio ordering.

The first kind of the stochastic ordering is called ordinary stochastic ordering. Suppose there are two random variables \( X \) and \( Y \) with distribution functions \( F(x) \) and \( G(x) \) and pdfs(or pmfs) \( f(x) \) and \( g(x) \), respectively. Then \( X \) is ordinary stochastically greater than \( Y \) if

\[
F(x) \leq G(x) \quad \text{for all } x,
\]

which is denoted by \( X \geq_{st} Y \).

The second kind of stochastic ordering is called uniformly stochastic ordering. \( X \geq_{ust} Y \) is defined as

\[
\frac{\tilde{F}(x)}{\tilde{G}(x)} \text{ is nondecreasing in } x,
\]

where \( \tilde{F}(x) \) and \( \tilde{G}(x) \) are the survival functions corresponding to \( F \) and \( G \), i.e. \( F = 1 - F \) and \( \tilde{G} = 1 - G \). This ordering is stronger than ordinary stochastic ordering, which means that uniformly stochastic ordering implies ordinary stochastic ordering.

The last ordering, which is the strongest ordering, is called likelihood ratio ordering. The condition required for likelihood ratio ordering is

\[
f(x)/g(x) \text{ is nondecreasing in } x,
\]
denoted as $X \overset{lr}{>} Y$. This ordering relationship implies the other two ordering.

It is well known that order statistics for continuous distribution satisfy likelihood ratio ordering, which implies the other two stochastic ordering. But in the literature, there is not any discussions related to uniformly stochastic ordering or likelihood ratio ordering of the order statistics from a discrete distribution. Here, we prove that the order statistics from a discrete distribution satisfy uniformly stochastic ordering.

**Theorem 5.1** Suppose $X_1, \ldots, X_H$ are i.i.d random variables having a common discrete cdf $F$. Let $X_{[r]}$, $X_{[r+1]}$ be the $r$, $r+1$-th order statistics with the cdf $F_{[r]}$, $F_{[r+1]}$. Then $\frac{1-F_{[r+1]}(x)}{1-F_{[r]}(x)}$ is a nondecreasing function in $x$.

**Proof:** From Arnold et al. (2008), we have

$$1 - F_{[k]}(x) = \sum_{i=0}^{k-1} \binom{H}{i} \{F(x)\}^i \{1 - F(x)\}^{H-i}, \quad k = r, r + 1$$

Then we can write $\frac{1-F_{[r+1]}(x)}{1-F_{[r]}(x)}$ as

$$\frac{1 - F_{[r+1]}(x)}{1 - F_{[r]}(x)} = 1 + \frac{\binom{H}{r} \{F(x)\}^r \{1 - F(x)\}^{H-r}}{\sum_{i=0}^{r-1} \binom{H}{i} \{F(x)\}^i \{1 - F(x)\}^{H-i}}.$$  \hspace{1cm} (5.1)

From (5.1), it is simple to see $\frac{1-F_{[r+1]}(x)}{1-F_{[r]}(x)}$ is a nondecreasing function in $x$.

From Theorem 5.1, by induction, it is easy to show $\frac{1-F_{[s]}(x)}{1-F_{[r]}(x)}$ for any $s > r$. Therefore, $F_{[1]}, \ldots, F_{[H]}$ satisfy uniformly stochastic ordering, i.e.

$$F_{[1]} \overset{ust}{<} F_{[2]} \overset{ust}{<} \ldots \overset{ust}{<} F_{[H]}.$$
5.3. Estimators under Ordinary Stochastic Ordering

In this section, we will introduce how to estimate the CDF of each judgment class under ordinary stochastic ordering. In the absence of any kind of stochastic ordering, the most common used estimator for \( r \)-th stratum CDF \( F_{[r]} \) is the empirical estimator, which is introduced in Chapter 2, i.e.

\[
\hat{F}_{[r]}^e(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_{[r,i]} \leq x\}.
\]

\( \hat{F}_{[r]}^e(x) \) is an unbiased estimator of \( F_{[r]} \) and the variance of \( \hat{F}_{[r]}^e(x) \) is

\[
Var(\hat{F}_{[r]}^e) = \frac{1}{n} F_{[r]}(1 - F_{[r]}).
\]

The EDF’s are also the maximum likelihood estimators of the strata CDF’s. This can be easily proven by writing the likelihood function of a ranked-set sample as

\[
\prod_{r=1}^{H} \prod_{j=1}^{k} C_{rj}(F_{[r]}(x_j) - F_{[r]}(x_{j-}))^{d_{rj}},
\]

where \( x_1 < ... < x_k \) are the \( k \) unique measured values and \( d_{rj} \) is the number of times that \( x_j \) is observed as the \( r \)-th order statistic in the sample.

The EDF’s, however, may violate the ordinary stochastic ordering constraint

\[
F_{[1]}(x) \geq ... \geq F_{[H]}(x)
\]

due to the randomness of the sampling procedures. Ozturk (2007) estimated the CDF of each order stratum from a continuous ranked set sample (both balanced and unbalanced) under ordinary stochastic ordering. Their method is to construct estimators of \( \{F_{[r]}, r = 1, \cdots, H\} \) which not only satisfy the ordinary stochastic ordering constraint (5.3) but also minimize the distance between the new estimates and the EDF’s. One option for the distance is given
by the weighted least square

\[
\sum_{r=1}^{H} \left\{ \hat{F}_r^e(x) - F_r(x) \right\}^2 n_r,
\]

where \( n_r \) is the sample size from \( r \)-th order stratum. The resulting estimators are called the isotonic regression estimators, denoted by \( \hat{F}_{iso}^r(x) \). There exist two analytical forms of \( \hat{F}_{iso}^r(x) \), given by

\[
\hat{F}_{iso}^r(x) = \min_{s \leq r} \max_{t \geq r} \sum_{g=s}^{t} n_g \frac{\hat{F}_g^e(x)}{n_{st}},
\]

or

\[
\hat{F}_{iso}^r(x) = \max_{t \geq r} \min_{s \leq r} \sum_{g=s}^{t} n_g \frac{\hat{F}_g^e(x)}{n_{st}},
\]

where \( n_{st} = \sum_{g=s}^{t} n_g \). Ozturk (2007) showed both formulas agree for a ranked set sample (either balanced or unbalanced). Numerically, the solutions can be obtained by using the pooled adjacent violators algorithm (PAVA). They showed by simulations that \( \hat{F}_{iso}^r(x) \) achieved a reduction on mean integrated square errors compared to the EDF’s in the context of continuous RSS.

### 5.4. Estimating Under Uniformly Stochastic Ordering

In this section, we introduce a new estimator of \( F_r \) by imposing the more stringent uniformly stochastic ordering constraint. The likelihood function of a ranked-set sample from a discrete distribution is shown in (5.2). Our goal is to find a set of estimators of \( F_r, r = 1, \ldots, H \) such that the uniform stochastic ordering constraint is satisfied. Maximizing (5.2) is equivalent to maximizing

\[
\prod_{r=1}^{H} \prod_{j=1}^{k} C_{rj} (\bar{F}_r(x_{j-1}) - \bar{F}_r(x_j))^{d_{rj}},
\]

which is

\[
\prod_{r=1}^{H} \prod_{j=1}^{k} C_{rj} \{1 - \frac{\bar{F}_r(x_j)}{\bar{F}_r(x_{j-1})}\}^{d_{rj}} \bar{F}_r(x_{j-1})^{d_{rj}}.
\]
By the definition of uniformly stochastic ordering, the constraint $F_{[1]} \overset{ust}{\leq} F_{[2]} \overset{ust}{\leq} \ldots \overset{ust}{\leq} F_{[H]}$ is equivalent to

$$\frac{F_{[r]}(x_j)}{F_{[r]}(x_{j-1})} \leq \frac{F_{[r+1]}(x_j)}{F_{[r+1]}(x_{j-1})}, \quad (5.6)$$

Now we can rewrite our problem of maximizing (5.6) subject to the constraint as follows. Let $\theta_{rj} = \frac{\bar{F}_{[r]}(x_j)}{\bar{F}_{[r]}(x_{j-1})}$ and $\bar{F}_{[v]}(x_0) = 1$. Then we have $\bar{F}_{[v]}(x_j) = \prod_{i=1}^{j} \theta_{ri}$ and (5.5) becomes

$$\prod_{r=1}^{H} \prod_{j=1}^{k} \theta_{rj}^{n_{rj}} (1 - \theta_{rj})^{d_{rj}},$$

where $n_{rj} = \sum_{k=j}^{H} d_{rk}$. The constraint becomes

$$\theta_{1j} \leq \theta_{2j} \leq \ldots \leq \theta_{Hj} \text{ for all } j.$$

The solution for this problem is given in Dykstra et al. (1991). It is the isotonic regression of the vector $(\hat{\theta}_{1j}, \ldots, \hat{\theta}_{Hj})$ with weights $(n_{1j}, \ldots, n_{Hj})$, where

$$\hat{\theta}_{rj} = \frac{n_{rj} - d_{rj}}{n_{rj}}.$$

After obtaining the solution, say $(\hat{\theta}_{1j}^*, \ldots, \hat{\theta}_{Hj}^*)$, we can compute $\hat{F}_{[r]}^{uso}(x_j)$ as

$$\hat{F}_{[r]}^{uso}(x_j) = 1 - \prod_{k=1}^{j} \theta_{rj}^*.$$

5.5. Simulation Studies

In this section, we conduct a simulation study to compare the performance of the EDF $\hat{F}_{[r]}^{e}$, Ozturk’s estimator $\hat{F}_{[r]}^{iso}$, and our new estimator $\hat{F}_{[r]}^{uso}$ for each order stratum. In our simulation studies, we set the set size ($H$) to be 3 and 5. For the underlying distributions, we chose the discrete uniform distribution ($N = 5, 10$), binomial distribution ($10, p = .2, .5$) and Poisson distribution ($\lambda = 5, 10$).
To compare the performance of those estimators, we define the relative efficiency of $\hat{F}_{i[r]}^i (i = iso, uso)$ as the ratio of mean integrated square errors (MISE), i.e.

$$RE(\hat{F}_{i[r]}^i) = \frac{MISE(\hat{F}_{i[r]}^e)}{MISE(\hat{F}_{i[r]}^i)}.$$  \hfill \text{(5.7)}

RE is estimated via 10000 replications for each combination of distribution parameter(s), set size and cycle size.

Figures 5.1, 5.2, and 5.3 show the simulated RE’s of $\hat{F}_{i[r]}^{iso}$ and $\hat{F}_{i[r]}^{uso}$ when the underlying distributions are discrete uniform distribution ($N_0$), binomial distribution, and Poisson distribution. The $x$-axis is the sample size from each order stratum. First, $\hat{F}_{i[r]}^{uso}$ usually outperforms $\hat{F}_{i[r]}^{iso}$ for those middle order strata (the strata except 1-st and $H$-th stratum). But for the edge strata (1-st and $H$-th stratum), especially $H$-th stratum, the performance of $\hat{F}_{i[r]}^{uso}$ may be worse than $\hat{F}_{i[r]}^{iso}$. Moreover, the advantages of $\hat{F}_{i[r]}^{uso}$ and $\hat{F}_{i[r]}^{iso}$ over $\hat{F}_{i[r]}^{e}$ diminish as the sample size of each stratum increases. Therefore, $\hat{F}_{i[r]}^{iso}$ and $\hat{F}_{i[r]}^{uso}$ are recommended for relative small sample size. Last, the RE’s of $\hat{F}_{i[r]}^{uso}$ and $\hat{F}_{i[r]}^{iso}$ are greater than 1 for each combination of parameters, which suggests $\hat{F}_{i[r]}^{uso}$ and $\hat{F}_{i[r]}^{iso}$ outperform $\hat{F}_{i[r]}^{e}$. 

60
Figure 5.1. Relative efficiency of $\tilde{F}_i^N(i = iso, uso)$ for discrete uniform distribution ($N_0$). The number $r$ on each line represents $r$-th order stratum.
Figure 5.2. Relative efficiency of $\hat{F}^i_r (i = \text{iso, uso})$ for binomial($n = 10, p$) distribution. The number $r$ on each line represents $r$-th order stratum.

Figure 5.3. Relative efficiency of $\hat{F}^i_r (i = \text{iso, uso})$ for Poisson distribution ($\lambda$). The number $r$ on each line represents $r$-th order stratum.
5.6. Conclusions

In this chapter, we proposed a new method to estimate the CDF of each stratum under discrete RSS. The new method was obtained by making use of the uniformly stochastic ordering restriction of judgment class distribution functions. It has been shown via simulation studies that the new method improves the relative efficiency over its competitors EDF $\hat{F}_{cr}$ and the estimator under ordinary stochastic ordering $\hat{F}_{iso}$ for the middle order strata (strata except the smallest stratum and the largest stratum). For the edge strata, especially the largest stratum, the new method may be worse than its competitors. Therefore, further study how to improve the estimation on edge strata is needed.
CHAPTER 6
DISCUSSION AND FUTURE DIRECTIONS

In this paper, we have discussed the ranked set sampling procedures when the underlying distribution is discrete. The difference between discrete RSS and continuous RSS is that ties may exist in discrete RSS. We have demonstrated that we can improve the estimation efficiency by using the information about ties. All procedures and conclusions are based on an assumption that ranking process is error-free, which may not be true in the real world. Therefore, ranking error models are needed for modeling the ranking process for discrete random variables more accurately.

6.1. Ranking Error Models

There are many ranking error models for continuous RSS in the literature. One popular model is to use a contaminated variable for ranking purposes. For example, the contaminated variable $Y$ for $X$ is defined as $Y = X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Then the ranking procedure is based on the value of $Y$. This ranking error model can be applied in discrete RSS with moderate modifications. For example, one possible modification is that

$$Y = X + \lfloor \epsilon \rfloor,$$

where $\epsilon \sim N(0, \sigma^2)$ and $\lfloor \epsilon \rfloor$ is the greatest integer which is less than $\epsilon$.

Another common used ranking error models in continuous RSS was proposed by Bohn and Wolfe (1994). The idea is to specify a $H \times H$ classification matrix $[p_{ij}]$, where $p_{ij}$ is probability that the unit that actually has rank $i$ is chosen as the $j$-th judgment order statistic. In Bohn and Wolfe (1994), one plausible configuration of $[p_{ij}]$ is given by the expected spacings model. In the expected spacings model, $p_{ij}$ is negatively proportional to
the difference of the expected value of \(i\)-th and \(j\)-th order statistic. This ranking error model can’t be applied in discrete RSS unless we have another model which models when ties are declared between different order strata. Adding another model, however, will increase the model complexity which may limit the use of the model.

6.2. Variations of Ranked Set Sampling

Another possible direction is to extend the methods proposed in this paper to the variations of ranked set sampling, for example, we can get the NPMLE of the CDF under judgment post-stratification (JPS). As mentioned in Chapter 1.2, JPS switches the measuring procedure and the ranking procedure. Although the procedures in JPS are different from those in RSS, the notations and the likelihood function obtained under two sampling schemes are same. Firstly, we can write a JPS sample as \(\{X_{[r]i}, l_{[r]i}, t_{[r]i}\}\), where \(l_{[r]i}\) is the number of units which are judged less than \(X_{[r]i}\) in the set which is used to rank \(X_{[r]i}\) and \(t_{[r]i}\) is the number of units which are judged tied to \(X_{[r]i}\) (including \(X_{[r]i}\) itself) in the set which is used to assess \(X_{[r]i}\). Similar to Chapter 2.3, we can define the corresponding \(l_j, t_j, v_j\) for all measured values \(x_j\) \((j = 1, \ldots, k)\). Then we can get the likelihood function of the JPS sample which is same to (2.8). By applying the same algorithm, we can obtain the approximate NPMLE of the CDF. The proposed NPMLE can also extend the application of the JPS. One important assumption of JPS in MacEachern et al. (2004) is that the units from each order stratum should be included at least once. Wang et al. (2012) considered the estimation problem of the CDF when there are empty strata. The NPMLE algorithm provides an alternative solution to this problem as the NPMLE doesn’t have any requirement on the sample size from each other stratum.
APPENDIX A
DISCRETE ORDER STATISTICS

Let $X$ be a random variable with probability density (or mass) function (pdf, or pmf) $f(x)$ and cumulative density function (cdf) $F(x)$. Let $X_r$ be $r$-th order statistic among $H$ units with probability density (or mass) function (pdf, or pmf) $f_r(x)$ and cumulative density function (cdf) $F_r(x)$. For a continuous r.v., the density function of $X_r$ is

$$f_r(x) = \frac{H!}{(r - 1)! (H - r)!} \{F(x)\}^{r-1} \{1 - F(x)\}^{H-r} f(x).$$

But for a discrete r.v., the pmf of $X_r$ has a different expression, which is

$$f_r(x) = \sum_{i=r}^{H} \binom{H}{i} \{F(x)\}^i [1 - F(x)]^{H-i} - \{F(x-\varepsilon)\}^i [1 - F(x-\varepsilon)]^{H-i}, \quad (A.1)$$

where $F(x-\varepsilon) = Pr(X < x)$.

(A.1) is known as the binomial sum form. There are also two other useful expressions of $f_r(x)$. One of them is called beta integral form which has the same expression for both continuous and discrete order statistics; that is

$$f_r(x) = \frac{H!}{(r - 1)! (H - r)!} \int_{F(x-\varepsilon)}^{F(x)} u^{r-1} (1 - u)^{H-r} du. \quad (A.2)$$

The other expression of $f_r(x)$ is derived by using a multinomial argument. It is given by

$$f_r(x) = \sum_{l=0}^{r-1} \sum_{u=0}^{H-l-u} \frac{H!}{(H - l - u)! (r - l - 1)! (u + r + 1)!} \{F(x-\varepsilon)\}^{r-1-l} \{1 - F(x)\}^{H-l-u} \{f(x)\}^{l+u+1}. \quad (A.3)$$

For more details, see Arnold et al. (2008).
Let \( l \) be the number of units which are smaller than \( r \)-th order statistic, \( t \) be the number of units which are tied to \( r \)-th order statistic, and \( H - l - t \) be the number of units which are greater than \( r \)-th order statistic. Then an equivalent form of (A.3) is

\[
f_{[r]}(x) = \sum_{l=0}^{r-1} \sum_{t=r-l-1}^{H-l-1} f_{[r]}(x, l, t) \tag{A.4}
\]

where

\[
f_{[r]}(x, l, t) = \frac{H! \{F(x-)\}^l \{1 - F(x)\}^{H-l-t} \{f(x)\}^t}{l! t! (H - l - t)!} \tag{A.5}
\]

Besides the difference on the expressions of \( f_{[r]} \), another difference between continuous order statistics and discrete order statistics is the Markov property. It is well known that order statistics from continuous distributions form a Markov chain, i.e.

\[
Pr(X_{[r+1]} = x | X_{[r]} = y, X_{[r-1]} = z) = Pr(X_{[r+1]} = x | X_{[r]} = y).
\]

But order statistics from a discrete distribution with at least three points in the support fail to form a Markov chain. A more detailed conclusion by Arnold et al. (2008) (page. 48) is

\[
Pr(X_{[r+1]} = x | X_{[r]} = y, X_{[r-1]} = z) < Pr(X_{[r+1]} = x | X_{[r]} = y). \tag{A.6}
\]

where \( x > y > z \) are elements in the support of the parent distribution. A direct conclusion from (A.6) is

\[
Pr(X_{[r+1]} = y | X_{[r]} = y, X_{[r-1]} = z) > Pr(X_{[r+1]} = y | X_{[r]} = y), z < y \tag{A.7}
\]

which suggests that the probability of getting a tie between \( r \)-th and \( r+1 \)-th order statistics depends on the value of \( r - 1 \)-th order statistic. Let’s introduce more results about ties among discrete order statistics.

**Theorem A.1** Let \( X_{[r]} \) and \( X_{[s]} \) be \( r \)-th and \( s \)-th order statistics among \( H \) units from a discrete distribution with pmf \( f(x) \) and cdf \( F(x) \). Let \( S \) be the support of the underlying
distribution. The probability that $r$-th and $s$-th ($r < s$) order statistics are tied is given by

$$Pr(X_r = X_s) = \sum_{x \in S} Pr(X_r = X_s = x)$$

where

$$Pr(X_r = X_s = x) = \frac{H!}{(H-s)! (s-r-1)! (r-1)!} \int_B u^{r-1}(v-u)^{s-r-1}(1-v)^{H-s}dvdu$$

(A.8)

and $B$ is given by $B = \{(u,v) : u \leq v, F(x-) \leq u, v \leq F(x)\}$.

**Proof:** By Theorem 3.3.1 in Arnold et al. (2008), we have the joint pmf of $X_r$ and $X_s$ which is

$$f_{r,s}(x,y) = Pr(X_r = x, X_s = y)$$

$$= \frac{H!}{(H-s)! (s-r-1)! (r-1)!} \int_B u^{r-1}(v-u)^{s-r-1}(1-v)^{H-s}dudv$$

where $x \leq y$ and $B = \{(u,v) : u \leq v, F(x-) \leq u \leq F(x), F(y-) \leq v \leq F(y)\}$. Set $y = x$ and then obtain (A.8).

A direct result from Theorem A.1 is

**Result 1** For $1 \leq r < H$, the probability that $X_r$ and $X_{r+1}$ are tied is given by

$$Pr(X_r = X_{r+1}) = \sum_{x \in S} Pr(X_r = X_{r+1} = x)$$

where

$$Pr(X_r = X_{r+1} = x) = H \binom{H-1}{r} \{I_F(x)(r+1, n-r) - I_F(x-)(r+1, n-r)\} -$$

$$- \binom{H}{r} F^r(x-) * \{(1 - F(x-))^{H-r} - (1 - F(x))^{H-r}\}$$

(A.9)

and $I_F(x)(i + 1, n - i)$ is Incomplete Beta function.
Proof: From (A.8), we have

\[
Pr(X_r = X_{r+1} = x) = \frac{H!}{(H-r-1)! (r-1)!} \int_{F(x)}^{F(x)} \int_{F(x)}^{u} u^{r-1} (1-v)^{H-r-1} dudv
\]

\[= H \binom{H-1}{r} \int_{F(x)}^{F(x)} v^r (1-v)^{H-r-1} dv \]

\[= - (H-r) \binom{H}{r} F^r(x) \int_{F(x)}^{F(x)} (1-v)^{H-r-1} dv \]

\[= H \binom{H-1}{r} \{I_{F(x)}(r+1, n-i) - I_{F(x)}(r+1, n-r)} \]

\[= - \binom{H}{r} F^r(x) \{ (1 - F(x))^{H-r} - (1 - F(x))^{H-r} \}. \]

Theorem A.1 suggests an easier way to compute the probability that \(r\)-th and \(s\)-th order statistics are tied given that the pmf of parent distribution is known. It is also interesting to investigate the expected number of ties among \(H\) units from a discrete distribution, say \(T_H\). Liu (2016) pointed out \(T_H\) is given by

\[T_H = \sum_{r=1}^{H-1} P(X_r = X_{r+1}). \]

By using Result 1, we can obtain an explicit expression of \(T_H\) by summing over \(r\). An easier way to compute \(T_H\) is given by noticing the relationship between \(T_H\) and \(T_{H+1}\).

Lemma A.1 Let \(T_H\) and \(T_{H+1}\) denote the expected number of ties among \(H\) and \(H+1\) units from a discrete distribution with pmf \(f(x)\). Then we have

\[T_{H+1} = T_H + \sum_{x \in S} f(x) \{1 - (1 - f(x))^H\}. \]

Proof: The proof is quite intuitive. \(H+1\) units can be grouped into a group of \(H\) units and a group of one unit. Then \(T_H\) is the expected number of ties in the first group. Moreover, the probability that the unit in the second group is tied to some units in the first group is \(1 - (1 - f(x))^H\). Therefore, the expected increase on ties number is given by
\[
\sum_{x \in S} f(x) \{1 - (1 - f(x))^H\}.
\]

Therefore we can obtain a explicit expression of \( T_H \).

**Theorem A.2** For any \( H \geq 1 \), \( T_H \) is given by

\[
T_H = H + \sum_{x \in S} \{(1 - f(x))^H - 1\}.
\]

**Proof:** Obviously, we have \( T_1 = 0 \). By induction using Lemma A.1, it is straightforward to obtain the expression of \( T_H \).

In Table A.1, we provide the expected number of ties among \( H = 2, 3, 4, 5 \) units with some common discrete distributions.

### Table A.1. Expected number of ties among \( H \) units for different underlying distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( H=2 )</th>
<th>( H=3 )</th>
<th>( H=4 )</th>
<th>( H=5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete Uniform (5)</td>
<td>0.200</td>
<td>0.560</td>
<td>1.048</td>
<td>1.638</td>
</tr>
<tr>
<td>Discrete Uniform (10)</td>
<td>0.100</td>
<td>0.290</td>
<td>0.561</td>
<td>0.905</td>
</tr>
<tr>
<td>Discrete Uniform (20)</td>
<td>0.050</td>
<td>0.148</td>
<td>0.290</td>
<td>0.476</td>
</tr>
<tr>
<td>Binomial(10, 0.1)</td>
<td>0.313</td>
<td>0.830</td>
<td>1.482</td>
<td>2.226</td>
</tr>
<tr>
<td>Binomial(10, 0.5)</td>
<td>0.176</td>
<td>0.493</td>
<td>0.923</td>
<td>1.443</td>
</tr>
<tr>
<td>Binomial(10, 0.7)</td>
<td>0.193</td>
<td>0.537</td>
<td>0.999</td>
<td>1.553</td>
</tr>
<tr>
<td>Binomial(20, 0.1)</td>
<td>0.214</td>
<td>0.591</td>
<td>1.091</td>
<td>1.685</td>
</tr>
<tr>
<td>Binomial(20, 0.5)</td>
<td>0.125</td>
<td>0.358</td>
<td>0.683</td>
<td>1.086</td>
</tr>
<tr>
<td>Binomial(20, 0.7)</td>
<td>0.137</td>
<td>0.390</td>
<td>0.740</td>
<td>1.173</td>
</tr>
<tr>
<td>Poisson(1)</td>
<td>0.309</td>
<td>0.820</td>
<td>1.465</td>
<td>2.200</td>
</tr>
<tr>
<td>Poisson(3)</td>
<td>0.167</td>
<td>0.468</td>
<td>0.879</td>
<td>1.379</td>
</tr>
<tr>
<td>Poisson(10)</td>
<td>0.090</td>
<td>0.260</td>
<td>0.503</td>
<td>0.810</td>
</tr>
</tbody>
</table>

For discrete distributions with a fix number of support values, the minimum and maximum of expected number of ties can be found.
**Theorem A.3** For a discrete distribution with finite support, say \( \{x_1, \cdots, x_N\} \), the minimum of \( T_H \) is achieved by the discrete uniform distribution for any \( H \geq 2 \), i.e.

\[
f(x_i) = \frac{1}{N}, i = 1, \cdots, N.
\]

and

\[
T_{H_{\text{min}}} = H - N(1 - (1 - \frac{1}{N})^H).
\]  

(A.10)

The maximum of \( T_H \) is achieved by any degenerated distribution, i.e.

\[
f(x_i) = \begin{cases} 
1, & \text{for one fixed } i \\
0, & \text{else}
\end{cases}
\]

and the maximum of \( T_H \) is

\[T_{H_{\text{max}}} = H - 1.\]

**Proof:** The maximum of \( T_H \) is straightforward.

Let \( p_i = f(x_i) \). To get the minimum, define the Lagrangian as

\[
\mathcal{L} = H + \sum_{i=1}^{N} \{(1 - p_i)^H - 1\} - \lambda(\sum_{i=1}^{N} p_i - 1).
\]

By taking the partial derivatives with respect to \( p_i, \lambda \), we have

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial p_i} &= -H(1 - p_i)^{H-1} - \lambda \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{i=1}^{N} p_i - 1
\end{align*}
\]

By setting them to 0 and solving those equations, we have \( p_i = \frac{1}{N}, i = 1, \cdots, N \). By plugging \( p_i \) into the expression of \( T_H \), we can obtain (A.10).
BIBLIOGRAPHY


