5-1-2019

K-Tuple Sampling From Partially Rank-Ordered Sets

Marvin Javier

K-TUPLE SAMPLING FROM PARTIALLY RANK-ORDERED SETS

by

Marvin Chris Javier

Bachelor of Arts in Mathematics and Molecular and Cell Biology
University of California, Berkeley
Dec. 2007

Master of Science in Statistics
San Diego State University
May 2013

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Mathematical Sciences

Department of Mathematical Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
May 2019

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

April 11, 2019

This dissertation prepared by

Marvin Chris Javier

entitled

K-Tuple Sampling from Partially Rank-Ordered Sets

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Mathematical Sciences
Department of Mathematical Sciences

Kaushik Ghosh, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Dean

Amei Amei, Ph.D.
Examination Committee Member

Hokwon Cho, Ph.D.
Examination Committee Member

Guogen Shan, Ph.D.
Graduate College Faculty Representative

ABSTRACT

K-TUPLE SAMPLING FROM PARTIALLY RANK-ORDERED SETS

by

Marvin Chris Javier

Dr. Kaushik Ghosh, Examination Committee Chair
University of Nevada, Las Vegas, USA

With the introduction of Ranked Set Sampling (RSS), McIntyre (1952) demonstrated that using ranking information to select units for measurement can lead to estimators with reduced variance when compared to their counterparts based on a simple random sample of the same size. This is done by selecting a set of units, and without direct measurement, ranking the units in the set before identifying one unit for measurement. This ranking of the units can be done through judgement ranking (such as visual assessment), or by using a correlated auxiliary variable.

In its original form, RSS does not allow for ties when ranking and requires a screening pool of size n^2. Several authors have tried to address these issues separately. In particular, Ghosh and Tiwari (2008) introduced k-Tuple Ranked Set Sampling, in which k measurements are made on each ranked set, thereby reducing the screening burden. Ozturk (2011) introduced Partially Rank-Ordered Set Sampling (PRSS) and showed that even when ties are allowed, estimators can still have lower variance when compared to their simple random sampling counterparts. Thus, the ranking burden can be reduced while still providing more efficient estimators.

In this dissertation, we generalize Ozturk's PRSS through application of Ghosh and

Tiwari's idea of k-tuple sampling, thus addressing both screening pool size reduction as well as ranking requirements. Named k-tuple Partial Rank-Ordered Set Sampling (KPRSS), three different sampling plans are presented: Uniform KPRSS, Balanced KPRSS, and General KPRSS. Partial Rank-Ordered Set Sampling, and by extension, Ranked Set Sampling, are special cases of KPRSS.

For Uniform and Balanced KPRSS, unbiased estimators of the population mean, variance, and distribution function are derived. It is shown that the variance of the sample mean and the variance of the empirical distribution function for these sampling plans are less than or equal to the variance of their simple random sample-based counterparts. Simulation studies as well as analysis of a data set of tree heights from Platt et al. (1988) are used to illustrate these results. For General KPRSS, an estimator of the population distribution function is derived along with its asymptotic properties.

Keywords: Ranked Set Sampling, Partially Rank-Ordered Set Sampling, k-Tuple Ranked Set Sampling, k-Tuple Partial Rank-Ordered Set Sampling

ACKNOWLEDGEMENTS

First I must thank my adviser, Professor Kaushik Ghosh, for his patience and guidance. He, along with Dr. Hokwon Cho, Dr. Amei Amei, and Dr. Guogen Shan, have contributed greatly to my knowledge and development. You are all wonderful professors and mentors.

I must also thank Dr. Petros Hadjicostas. His help and insight showed me that Part 3 of Theorem 2.4 could be proved.

Last, but certainly not least, I thank my wife, Lyndsie Javier. Her support keeps me motivated even in the most difficult times. My success is because of her love and motivation.

DEDICATION

To Lyndsie Javier

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1:

INTRODUCTION

An important, desirable property in statistical inference is that the sample be an accurate reflection of the population. Gathering a representative sample is especially important when measurements are difficult to perform and/or are expensive. Simple random sampling (SRS) is the typical way of gathering a sample. Think of pulling names from a hat: everyone in the population has an equal chance of being selected and are picked randomly. However, when sampling this way, there is a chance that we gather an extreme sample. For example, gathering a sample of people all over 6 feet tall when studying average height.

First introduced by McIntyre (1952), ranked set sampling (RSS) is a sampling design which takes advantage of ranking information to generate a sample that is more representative of the population than simple random sampling (SRS). To obtain an RSS of size n, one proceeds through the following steps:

1. Select a random sample of n units.

2. Rank the n units , thus forming a ranked set. Ranking is done through visual inspection (also called judgement ranking) or through the use of a correlated variable, but without direct measurement of the characteristic of interest.

3. Measure the characteristic of interest for the lowest ranked unit.

4. Repeat steps 1 and 2, and measure the second lowest ranked unit.

5. Repeat this procedure until the nth ranked unit is measured from the nth set.

For a sample of size n, this method draws from each of the n order statistics. This helps to ensure that the data selected includes observations from the entire range of the population. Dell and Clutter (1972) showed that under perfect ranking, the estimate of the mean based on RSS is unbiased and has smaller variance compared to the estimate based on an SRS of the same size.

An example of judgement ranking would be to visually line up people by height when the goal is to estimate average height. An example of using a correlated variable can be found in Chen et al. (2004), in which the authors present a case study for the yield of bark from Cinchona plants. Measurement of actual yield requires uprooting and drying of the bark from a plant, which is slow and expensive. However, bark yield is highly correlated to bark volume, which can be easily calculated from the height and girth of the plant. Thus bark volume can be used as the ranking variable.

Another example of using a correlated ranking variable is demonstrated by Yu and Lam (1997). They retrospectively applied ranked set sampling to the data from a 1975 study to estimate the amount of plutonium in surface soil around Nevada Test Site Area 13. The correlated variable used was "Field Instrument for Determination of Low Energy Radiation (FIDLER)". This reading could be taken in the field to identify potential areas for further investigation. Samples could then be taken back to the lab for radiochemical analysis; a process which was 50 times more expensive. Yu and Lam found the ranked set sampling

procedures showed significant improvement over simple random samples.

What these examples demonstrate is that ranked set sampling is extremely useful when ranking is done cheaply and more easily than measuring the variable of interest. Applications of this method of sampling can be found in public health (Chen et al., 2007, Ozturk and MacEachern, 2007), Ecology (Mode et al., 1999), genetic linkage analysis (Zheng et al., 2006) and manufacturing (Asghari et al., 2017). The trade-off for reduction in variance comes from screening n^2 units to obtain a sample of size n.

When the cost of gathering and ranking units becomes significant, the need for filtering through n^2 units may not be practical. One option to reduce this number is to reduce n, the size of each ranked set. For a sample of size 30, instead of gathering 30 simple random samples of size 30 each, one could gather a ranked set sample of size 6, and repeat the process 5 times. The first procedure screens 900 units, the second screens 180 ($= 6^2 \times 5$). Nahhas et al. (2002) developed a cost model which can be used to find the optimal set size, n, for a fixed cost.

In contrast, Muttlak (1996) introduced paired rank set sampling. Under the paired RSS, the i-th and $n - i + 1$-th elements are selected for measurement from the i-th ranked set. Paired RSS and it's generalization, hybrid RSS (Haq et al., 2016), still collect at most 2 measurements from each ranked set.

Wang et al. (2004) proposed general ranked set sampling and expanded past two measurements. Under this scheme τ measurements are made from each ranked set. All $\binom{n}{\tau}$ combinations are cycled through. Ghosh and Tiwari (2008) used the term balanced k-tuple ranked set sample (for the k observed units from each ranked set) for the procedure from Wang et al. (2004), and expanded the theory to unbalanced k-tuple ranked set sample.

Under this framework, they developed an estimator of the distribution function. One commonality to these plans is that the size of each k-tuple remains constant (i.e. the number of measurements from each ranked set are the same). Ghosh and Tiwari (2009) removed this requirement and used the results to develop two sample test for means, control percentile test, and Wilcoxon-Mann-Whitney test.

For ranked set sampling there is often the assumption of perfect ranking. But the variance of RSS estimators are affected by ranking errors. Dell and Clutter (1972) demonstrate this for the sample mean. The problem of imperfect rankings has been studied in multiple ways. Aragon et al. (1999) used a ranking error probability matrix to study the effect of ranking errors. Ozturk (2010) estimate within-set ranking errors using nonparametric maximum-likelihood. See Wolfe (2012) for an overview of other methods.

The simplest way to reduce ranking errors is to allow ties. Partially ranked-ordered set sampling (PROSS), proposed by Ozturk (2011), allows the inclusion of ties when ranking. Groups of tied units can then be ranked against other groups, but units within a group are considered equivalent. Ozturk proposes three PROSS sampling designs; the simplest of which is denoted by G^{**}. To obtain a G^{**} PROSS of size n, the steps are as follows:

1. Select a random sample of n units.

2. Form G groups of equal sizes and rank the groups. Thus, a partially rank-ordered set is formed.

3. From the first (lowest ranked) group, randomly select a unit for measurement.

4. Repeat steps 1-3, each time selecting a new random sample of n units and randomly

Group Size	$Var(\overline{X}_{RSS})$	$Var(\overline{X}_{PROSS})$	$Var(\overline{X}_{RSS})$ / $Var(\overline{X}_{PROSS})$
2	0.0176	0.0169	1.0439
3	0.0226	0.0222	1.0210
4	0.0284	0.0251	1.1312
6	0.0360	0.0354	1.0173

Table 1.1: $Var(\overline{X}_{RSS})$ / $Var(\overline{X}_{PROSS})$. Based on 1,000 replicates with intra-group ranking errors using samples of size 12 from the standard normal distribution.

selecting one unit for measurement from the next highest ranked group. Continue until a unit has been selected from a group of each rank.

5. Repeat steps 1-4 until n units have been measured.

The most general version of PROSS allows groups to be of different sizes, as well as the number of groups in step 2 to vary. When groups are of size one, PROSS becomes RSS. Ozturk (2011) developed estimators of the population mean and variance, and Nazari and Jozani (2014) developed an estimator of the distribution function from a PROSS sample. As mentioned before, by allowing ties, PROSS reduces ranking errors. When ranking errors occur, the variance of estimators from PROSS can be lower than their RSS-based counterpart.

Table 1.1 shows the result of a simulation study based on 1,000 replicates with total sample size of 12 from the standard normal distribution. Set size was 12 for both RSS and PROSS. Units within a group were randomly assigned a rank, but ranking between groups was perfect. For all group sizes, $Var(\overline{X}_{RSS}) > Var(\overline{X}_{PROSS})$. Notice, that as group sizes increased, so did the variance of $\overline{X}_{RSS}$. This is because a larger group size means that there is a greater change of selecting a miss-ranked unit when performing ranked set sampling.

Thinking of various sampling procedures as a spectrum, on one end of it is simple random

sampling which requires no ranking, has the smallest number of units screened (all units selected are measured), and gives rise to estimators with greater variance than ranked set sampling. On the other end is ranked set sampling, with the greatest ranking requirements, largest number of screened units, and the lowest variance. partially rank-ordered set sampling falls in between these two extremes in terms of ranking requirements and variance, yet still only measures one unit from each ranked set. k-tuple ranked set sampling address multiple measurements, yet still does not allow ties.

This dissertation looks to generalize partially rank-ordered set sampling to allow k measurements (just as k-tuple ranked set sampling generalizes ranked set sampling). By doing so, this method hopes to simultaneously address the problems of ranking error and screening pool size associated with RSS and further fill in the spectrum between simple random sampling and ranked set sampling. We will work with continuous distributions and assume perfect ranking between partially rank-ordered sets.

In Chapter 2, we introduce the most strict k-tuple partially rank-ordered set sampling (KPRSS) method; which we call Uniform KPRSS. In this scheme, all partially rank-ordered sets are of the same size and all ranked groups are of the same size. A measurement is then taken from each ranked group. Estimators of the sample mean, sample variance, and distribution function are derived.

Chapter 3 introduces Balanced KPRSS. This is a generalization of Uniform KPRSS. Here, the number of measurements from each partially rank-ordered set, k, is less than or equal to the number of groups. Again, estimators of the sample mean, sample variance, and distribution function are derived.

In Chapter 4, General KPRSS is introduced. For this scheme, partially rank-ordered

set size can change with each SRS. Groups need not be of the same size. Number of measurements for each partially rank-ordered set size, k, can also change. An estimator of the distribution function along with its asymptotic properties are derived.

This dissertation is concluded in Chapter 5. There, a summary of the presented work is ppresented along with possible future extensions.

CHAPTER 2:

UNIFORM K-TUPLE PARTIALLY RANK-ORDERED SET SAMPLING

In this chapter, we introduce the simplest of the k-tuple partially rank-ordered set sampling schemes. Under the assumption of perfect ranking of groups in each partially rank-ordered set, we derive unbiased estimators of the population mean, variance, and distribution function.

2.1 Sampling Procedure and Preliminary Results

For a uniform k-tuple partially rank-ordered set sample from a population with cdf $F(x)$ and density $f(x)$, assume that N, n, and g are pre-specified. To collect a sample of size N, using a set size of n, the proposed sampling scheme is as follows:

1. Randomly select n units.

2. Partially rank the n units into k groups, each of size g ($n = kg$). All items in the first group are considered smaller than those in the second group. All items in the second group are considered smaller than the third group, etc. However, no ordering is imposed on the units within a group. This forms a partially ranked-ordered set of size n. The ranking can be done using visual judgement, or measurement of a

characteristic related to the characteristic of interest, but without direct measurement of the characteristic of interest.

3. From each of the k groups, select a unit at random to be fully measured. These measurements constitute the k-tuple for that partially rank-ordered set.

4. Repeat M times until N $(= kM)$ units have been measured.

Denoted by UKPRSS(k, g, M), we call this a uniform k-tuple partially rank-ordered set sample. The number of groups and the sample size does not change for each partially rank-ordered set. Thus our sample can be regarded as having M independent realizations of a single k-tuple. We use the following notations:

1. $[a, b, n]s$: is the partially ordered group that contains the a-th through b-th order statistics from the s-th simple random sample of size n.

2. $X_{[a,b,n]s}$: is the measurement from $[a, b, n]s$. When the value of s is obvious or irrelevant, $X_{[a,b,n]}$ will be used.

The following lemma is from Ozturk (2011). It is presented here using the above notation.

Lemma 2.1. *(Ozturk, 2011)*

Let $f_{(i:n)}(x)$, $\mu_{(i:n)}$, and $\sigma^2_{(i:n)}$ be the density function, mean, and variance, respectively, of the i-th order statistic from a sample of size n. Then,

1. The density of $X_{[a,b,n]}$ is $f_{[a,b,n]}(x) = \dfrac{1}{b-a+1} \displaystyle\sum_{i=a}^{b} f_{(i:n)}(x)$

2. $E\left(X_{[a,b,n]}\right) = \dfrac{1}{b-a+1}\sum_{i=a}^{b}\mu_{(i:n)}$

3. $Var\left(X_{[a,b,n]}\right) = \dfrac{1}{b-a+1}\sum_{i=a}^{b}(\sigma^2_{(i:n)} + \mu^2_{(i:n)}) - \left(\dfrac{1}{b-a+1}\sum_{i=a}^{b}\mu_{(i:n)}\right)^2$

Building on Ozturk's result, we now derive other properties of measurements from partially rank-ordered sets.

Lemma 2.2. *The distribution function of* $X_{[a,b,n]}$ *is:*

$$F_{[a,b,n]}(x) = \frac{1}{b-a+1}\sum_{i=a}^{b}F_{(i:n)}(x),$$

where $F_{(i:n)}(x)$ *is the distribution function of the* i*-th order statistic from a sample of size* n*.*

Proof. The result follows from integration of the density function. □

Lemma 2.3. *For* $(a_c, b_c) \neq (a_d, b_d)$*, let* $g = b_c - a_c + 1 = b_d - a_d + 1$*. Then,*

$$E(X_{[a_c,b_c,n]s}X_{[a_d,b_d,n]s}) = \frac{1}{g^2}\sum_{i=a_c}^{b_c}\sum_{j=a_d}^{b_d}\mu_{(i,j)}$$

where $\mu_{(i,j)} = E\left(X_{(i:n)}X_{(j:n)}\right)$ *is the joint mean of the* i*-th and* j*-th order statistics.*

Proof. Let $X_{[a_c,b_c,n]s} \sim X_{(i:n)}$ denote the event that the measurement $X_{[a_c,b_c,n]s}$ was drawn from $X_{(i:n)}$, the ith order statistic, where $a_c \leq i \leq b_c$.

$$\begin{aligned}E(X_{[a_c,b_c,n]s}X_{[a_d,b_d,n]s}) =& E[E(X_{(i:n)}X_{(j:n)}|X_{[a_c,b_c,n]s} \sim X_{(i:n)}, X_{[a_d,b_d,n]s} \sim X_{(j:n)})]\\ =& \frac{1}{g^2}\sum_{i=a_c}^{b_c}\sum_{j=a_d}^{b_d}\mu_{(i,j)}\end{aligned}$$

□

Corollary 2.1. *For* $(a_c, b_c) \neq (a_d, b_d)$,

$$Cov(X_{[a_c,b_c,n]s}, X_{[a_d,b_d,n]s}) = \frac{1}{g^2} \sum_{i=a_c}^{b_c} \sum_{j=a_d}^{b_d} Cov(X_{(i:n)}, X_{(j:n)}) \tag{2.1.1}$$

Proof.

$$\begin{aligned} Cov(X_{[a_c,b_c,n]s}, X_{[a_d,b_d,n]s}) =& E(X_{[a_c,b_c,n]s} X_{[a_d,b_d,n]s}) - E(X_{[a_c,b_c,n]s}) E(X_{[a_d,b_d,n]s}) \\ =& \frac{1}{g^2} \sum_{i=a_c}^{b_c} \sum_{j=a_d}^{b_d} \mu_{(i,j)} - \frac{1}{g^2} \sum_{i=a_c}^{b_c} \sum_{j=a_d}^{b_d} \mu_{(i:n)} \mu_{(j:n)} \\ =& \frac{1}{g^2} \sum_{i=a_c}^{b_c} \sum_{j=a_d}^{b_d} Cov(X_{(i:n)}, X_{(j:n)}) \end{aligned}$$

□

2.2 Estimation of the Population Mean

Consider an UKPRSS(k, g, M) from a population with mean μ and variance σ^2. In this section we address estimation of the population mean. The natural estimator of the population mean based on a UKPRSS is the sample mean:

$$\overline{X}_{UKRPSS} = \frac{1}{kM} \sum_{s=1}^{M} \sum_{i=1}^{k} X_{[a_i,b_i,n]s}, \tag{2.2.1}$$

where, $a_i \equiv (i-1)g + 1, b_i \equiv (ig)$, and $i = 1, \ldots, k$.

Theorem 2.1. *Suppose a Uniform k-Tuple Partially Rank-Ordered Set Sample is drawn from a distribution* F *and density* f *which is continuous and strictly positive on* $\{x|0 < F(x) < 1\}$. *Then,*

1. $\overline{X}_{UKRPSS}$ *is an unbiased estimator of the population mean* μ.

2. The variance of $\overline{X}_{UKRPSS}$ *is*

$$\begin{aligned} Var(\overline{X}_{UKRPSS}) =& \frac{k}{n}\sigma^2_{\bar{X}_{SRS}} + \frac{1}{N}E\left(X^2\right) \\ & - \frac{k}{Nn^2}\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\leq j_1,j_2\leq b_i}\sigma_{(j_1,j_2:n)}\right], \end{aligned}$$

where:

- $g = b_i - a_i + 1\ \forall i$,
- $\sigma^2_{\bar{X}_{SRS}}$ *is the variance of the sample mean from an SRS of size* $N = kM$,
- $\sigma_{(j_1,j_2:n)}$ *is the covariance between the* j_1*-th and* j_2*-th order statistics from an SRS of size* n.

3. $Var(\overline{X}_{UKPRSS}) \leq Var(\overline{X}_{SRS})$.

Proof.

1. Unbiasedness

Let $X_{[a_j,b_j,n]s}$ be the jth element of the k-tuple from SRS s. Since all group sizes are the same, we have $b_j - a_j + 1 \equiv g\ \forall j$. Define

$$\overline{X}_s = \frac{1}{k}\sum_{j=1}^{k} X_{[a_j,b_j,n]s} = \frac{g}{n}\sum_{j=1}^{k} X_{[a_j,b_j,n]s}.$$

Then, $\overline{X}_{UKRPSS} = \frac{1}{M}\sum_{s=1}^{M}\overline{X}_s$ Now note that

$$\begin{aligned} E\overline{X}_s =& E\left(\frac{g}{n}\sum_{j=1}^{k} X_{[a_j,b_j,n]s}\right) \\ =& \frac{g}{n}\sum_{j=1}^{k} E(X_{[a_j,b_j,n]s}) \\ =& \frac{g}{n}\sum_{j=1}^{k}\frac{1}{g}\sum_{i=a_j}^{b_j}\mu_{(i:n)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{j=1}^{n}\mu_{(j:n)} \\
=&\frac{1}{n}n\mu \\
=&\mu.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
E\overline{X}_{UKRPSS} =&\frac{1}{M}\sum_{s=1}^{M}E\overline{X}_s \\
=&\frac{1}{M}\sum_{s=1}^{M}\mu \\
=&\mu.
\end{aligned}
$$

2. Variance

$$
\begin{aligned}
Var(\overline{X}_s) =&Var\left(\frac{g}{n}\sum_{j=1}^{k}X_{[a_j,b_j,n]s}\right) \\
=&\left(\frac{g}{n}\right)^2\left[\sum_{j=1}^{k}Var(X_{[a_j,b_j,n]s})+\sum_{c=1}^{k}\sum_{\substack{d=1\\d\neq c}}^{k}Cov(X_{[a_c,b_c,n]s},X_{[a_d,b_d,n]s})\right]. \qquad (2.2.2)
\end{aligned}
$$

Applying equation (2.1.1) into (2.2.2) gives:

$$
\begin{aligned}
Var(\overline{X}_s) =&\left(\frac{g}{n}\right)^2\left[\sum_{j=1}^{k}Var(X_{[a_j,b_j,n]s})+\sum_{c=1}^{k}\sum_{\substack{d=1\\d\neq c}}^{k}\frac{1}{g^2}\sum_{i=a_c}^{b_c}\sum_{j=a_d}^{b_d}Cov(X_{(i:n)},X_{(j:n)})\right] \\
=&\left(\frac{g}{n}\right)^2\left[\frac{1}{g}\sum_{i=1}^{n}E(X_{(i:n)}^2)-\frac{1}{g^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2\right. \\
&\left.+\sum_{c=1}^{k}\sum_{\substack{d=1\\d\neq c}}^{k}\frac{1}{g^2}\sum_{i=a_c}^{b_c}\sum_{j=a_d}^{b_d}Cov(X_{(i:n)},X_{(j:n)})\right] \\
=&\left(\frac{g}{n^2}\right)\sum_{i=1}^{n}E(X_{(i:n)}^2)-\frac{1}{n^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 \\
&+\frac{1}{n^2}\sum_{c=1}^{k}\sum_{\substack{d=1\\d\neq c}}^{k}\sum_{i=a_c}^{b_c}\sum_{j=a_d}^{b_d}Cov(X_{(i:n)},X_{(j:n)}).
\end{aligned}
$$

Notice that the last term is all inter-group covariances. This can be re-written as the sum of all covariances minus all intra-group covariances:

$$\sum_{c=1}^{k}\sum_{\substack{d=1\\d\neq c}}^{k}\sum_{i=a_c}^{b_c}\sum_{j=a_d}^{b_d} Cov(X_{(i:n)},X_{(j:n)}) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_{(i:n)},X_{(j:n)}) - \sum_{i=1}^{k}\sum_{a_i\leq j_1,j_2\leq b_i} Cov(X_{(j_1:n)},X_{(j_2:n)}).$$

Also note that

$$E\left(\sum_{i=1}^{n} X_{(i:n)}^2\right) = E\left(\sum_{i=1}^{n} X_i^2\right) = nE(X^2),$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_{(i:n)},X_{(j:n)}) = n\sigma^2.$$

This gives:

$$\begin{aligned} Var\left(\overline{X}_s\right) =& \frac{g}{n}E\left(X^2\right) - \frac{1}{n^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \frac{1}{n}\sigma^2 - \frac{1}{n^2}\sum_{i=1}^{k}\sum_{a_i\leq j_1,j_2\leq b_i} Cov(X_{(j_1:n)},X_{(j_2:n)}) \\ =& \frac{1}{kg}\sigma^2 + \frac{1}{k}E\left(X^2\right) - \frac{1}{n^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 - \frac{1}{n^2}\sum_{i=1}^{k}\sum_{a_i\leq j_1,j_2\leq b_i} Cov(X_{(j_1:n)},X_{(j_2:n)}). \end{aligned}$$

Then we have:

$$\begin{aligned} Var(\overline{X}_{UKRPSS}) =& Var\left(\frac{1}{M}\sum_{s=1}^{M} Var(\overline{X}_s)\right) \\ =& \frac{1}{M^2}(M)Var(\overline{X}_s) \\ =& \left(\frac{1}{M}\right)\left[\frac{1}{kg}\sigma^2 + \frac{1}{k}E\left(X^2\right) - \frac{1}{n^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2\right. \\ & \left. - \frac{1}{n^2}\sum_{i=1}^{k}\sum_{a_i\leq j_1,j_2\leq b_i} Cov(X_{(j_1:n)},X_{(j_2:n)})\right] \\ =& \frac{1}{g}\sigma^2_{\overline{X}_{SRS}} + \frac{1}{N}E\left(X^2\right) - \frac{1}{Mn^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 \end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{Mn^2}\sum_{i=1}^{k}\sum_{a_i\le j_1,j_2\le b_i} Cov(X_{(j_1:n)},X_{(j_2:n)})\\
=&\frac{k}{n}\sigma^2_{\overline{X}_{SRS}} + \frac{1}{N}E\left(X^2\right) - \frac{k}{Nn^2}\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2\\
&- \frac{k}{Nn^2}\sum_{i=1}^{k}\sum_{a_i\le j_1,j_2\le b_i} Cov(X_{(j_1:n)},X_{(j_2:n)}).
\end{aligned}
$$

3. Inequality

To have $Var(\overline{X}_{UKPRSS}) \le Var(\overline{X}_{SRS})$, we must show that

$$
\frac{1}{N}E\left(X^2\right) - \frac{k}{Nn^2}\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\le j_1,j_2\le b_i}\sigma_{(j_1,j_2:n)}\right] \le \frac{n-k}{n}\sigma^2_{\overline{X}_{SRS}}.
$$

This is true if and only if:

$$
\begin{aligned}
& E\left(X^2\right) - \frac{k}{n^2}\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\le j_1,j_2\le b_i}\sigma_{(j_1,j_2:n)}\right] \le \frac{n-k}{n}\sigma^2\\
\Leftrightarrow\quad & E\left(X^2\right) - \frac{k}{n^2}\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\le j_1,j_2\le b_i}\sigma_{(j_1,j_2:n)}\right] \le \frac{n-k}{n}\left(E\left(X^2\right)-\mu^2\right)\\
\Leftrightarrow\quad & \frac{k}{n^2}\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\le j_1,j_2\le b_i}\sigma_{(j_1,j_2:n)}\right] \ge \frac{k}{n}E\left(X^2\right) + \frac{n-k}{n}\mu^2\\
\Leftrightarrow\quad & k\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\le j_1,j_2\le b_i}\sigma_{(j_1,j_2:n)}\right] \ge nkE\left(X^2\right) + n(n-k)\mu^2\\
\Leftrightarrow\quad & k\sum_{i=1}^{k}\left[\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + \sum_{a_i\le j_1,j_2\le b_i}\sigma_{(j_1,j_2:n)}\right] \ge nk\sigma^2 + n^2\mu^2\\
\Leftrightarrow\quad & k\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + nk\sigma^2 + k\sum_{i=1}^{k}\sum_{\substack{a_i\le j_1,j_2\le b_i\\ j_1\ne j_2}}\sigma_{(j_1,j_2:n)} \ge nk\sigma^2 + n^2\mu^2 \qquad (2.2.3)\\
\Leftrightarrow\quad & k\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + k\sum_{i=1}^{k}\sum_{\substack{a_i\le j_1,j_2\le b_i\\ j_1\ne j_2}}\sigma_{(j_1,j_2:n)} \ge n^2\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2\\
\Leftrightarrow\quad & k\sum_{i=1}^{k}\left(\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2 + k\sum_{i=1}^{k}\sum_{\substack{a_i\le j_1,j_2\le b_i\\ j_1\ne j_2}}\sigma_{(j_1,j_2:n)} \ge \left(\sum_{i=1}^{k}\sum_{j=a_i}^{b_i}\mu_{(j:n)}\right)^2. \qquad (2.2.4)
\end{aligned}
$$

Inequality (2.2.3) is due to the fact that

$$\sum_{i=1}^{n} \sigma_{(i)}^2 = \sum_{i=1}^{n} \sigma_i^2 = n\sigma^2.$$

By the Cauchy-Schwarz inequality, $k \sum_{i=1}^{k} \left(\sum_{j=a_i}^{b_i} \mu_{(j:n)} \right)^2 \geq \left(\sum_{i=1}^{k} \sum_{j=a_i}^{b_i} \mu_{(j:n)} \right)^2$. Also, because the covariance of order statistics is always non-negative (Bickel, 1967), inequality (2.2.4) is always true.

□

The next section demonstrates the results of Theorem 2.1 with simulations from the standard normal and gamma distributions.

2.2.1 Simulation Results

Tables 2.1 and 2.2 show the comparison between $\overline{X}_{UKRPSS}$ and $\overline{X}_{SRS}$ for a total sample size of $N = 24$ for the standard normal distribution and gamma distribution with scale = 1 and shape = 2, respectively. Ratios of the theoretical variances are given in the last columns of each table. For the UKRPSS, each partially rank-ordered set was of size $n = 6$. Group sizes were adjusted accordingly for each value of k. Theoretical calculations were done using the covariance and expected values of order statistics from Harter and Balakrishnan (1996). Simulations were done with a total sample size of $N = 24$ and partially rank-ordered set size of $n = 6$. $\overline{X}_{UKPRSS}$ was then calculated for the sample. This process was repeated 10 million times and the sample average and standard deviation were used to estimate $\mu_{\overline{X}_{UKRPSS}}$ and $\sigma^2_{\overline{X}_{UKPRSS}}$.

k	Group Size	μ_{SRS} Simulated	μ_{UKRPSS} Simulated	Var($\overline{X}_{SRS}$)	Var($\overline{X}_{UKRPSS}$)	Var($\overline{X}_{UKRPSS}$) Simulated	$\frac{Var(\overline{X}_{SRS})}{Var(\overline{X}_{UKRPSS})}$
1	6	0.0001	0.0002	0.0417	0.0417	0.0417	1.0000
2	3	0.0004	-0.0001	0.0417	0.0256	0.0256	1.6274
3	2	-0.0003	0.0003	0.0417	0.0265	0.0264	1.5732
6	1	-0.0001	0.0002	0.0417	0.0417	0.0417	1.0000

Table 2.1: $\overline{X}_{UKPRSS}$ vs $\overline{X}_{SRS}$ for the standard normal distribution. Ratios in the last column are based on theoretical variances.

k	Group Size	μ_{SRS} Simulated	μ_{UKRPSS} Simulated	Var($\overline{X}_{SRS}$)	Var($\overline{X}_{UKRPSS}$)	Var($\overline{X}_{UKRPSS}$) Simulated	$\frac{Var(\overline{X}_{SRS})}{Var(\overline{X}_{UKRPSS})}$
1	6	1.99989485	1.99996734	0.08333333	0.083362957	0.083344842	0.99964464
2	3	2.00002886	1.99993594	0.08333333	0.054869335	0.054809241	1.518759676
3	2	2.00008053	1.9999622	0.08333333	0.055569403	0.055505521	1.499626207
6	1	2.00014846	2.00001977	0.08333333	0.083511022	0.083365602	0.997872269

Table 2.2: $\overline{X}_{UKPRSS}$ vs $\overline{X}_{SRS}$ for gamma distribution with scale = 1 and shape = 2

The theoretical and simulated values for $\sigma^2_{\overline{X}_{UKPRSS}}$ are in close agreement in both cases. As expected, the results show $\overline{X}_{UKPRSS}$ to be a more efficient estimator than $\overline{X}_{SRS}$ for estimating μ.

2.3 Estimation of the Population Variance

Next, we concentrate on estimation of the population variance. First, consider the sample variance given in equation (2.3.1).

Theorem 2.2. *The usual sample variance,*

$$S^2 = \frac{1}{N-1}\sum_{s=1}^{M}\sum_{i=1}^{k}(X_{[a_i,b_i,n]s} - \overline{X}_{UKPRSS})^2 \tag{2.3.1}$$

is a biased estimator of the population variance, σ^2.

Proof.

$$E(S^2) = \frac{1}{N-1}E\left[\sum_{s=1}^{M}\sum_{i=1}^{k}(X_{[a_i,b_i,n]s} - \overline{X}_{UKRPSS})^2\right]$$

$$
\begin{aligned}
=&\frac{1}{N-1}E\left[\sum_{s=1}^{M}\sum_{i=1}^{k}(X^2_{[a_i,b_i,n]s}-2\overline{X}_{UKRPSS}X_{[a_i,b_i,n]s}+\overline{X}^2_{UKRPSS})\right]\\
=&\frac{1}{N-1}\left[MkE(X^2)-MkE(\overline{X}^2_{UKRPSS})\right]\\
=&\frac{kM}{N-1}\left[\sigma^2+\mu^2-Var(\overline{X}_{UKRPSS})-\mu^2\right]\\
=&\frac{kM}{N-1}\left[\sigma^2-Var(\overline{X}_{UKRPSS})\right].
\end{aligned}
$$

□

Because of this, we follow the approach of Ozturk (2011) and MacEachern et al. (2002) and define an estimator of the population variance as follows:

$$
\begin{aligned}
\hat{\sigma}^2_{UKRPSS}=&\frac{1}{2M(M-1)n^2}\sum_{s_1\neq s_2}^{M}\left[\sum_{i=1}^{k}\left(gX_{[a_i,b_i,n]s_1}-gX_{[a_i,b_i,n]s_2}\right)^2\right.\\
&\left.+\sum_{i\neq j}^{k}\left(gX_{[a_i,b_i,n]s_1}-gX_{[a_j,b_j,n]s_2}\right)^2\right]
\end{aligned}
$$

Notice that $\hat{\sigma}^2_{UKRPSS}$ uses both the within-group and between-group variation.

Theorem 2.3.

$\hat{\sigma}^2_{UKRPSS}$ is an unbiased estimator for the population variance. That is,

$$E\left(\hat{\sigma}^2_{UKRPSS}\right)=\sigma^2$$

Proof. Let

$$A=\frac{1}{2M(M-1)n^2}\sum_{s_1\neq s_2}^{M}\sum_{i=1}^{k}\left(gX_{[a_i,b_i,n]s_1}-gX_{[a_i,b_i,n]s_2}\right)^2$$

and

$$B=\frac{1}{2M(M-1)n^2}\sum_{s_1\neq s_2}^{M}\sum_{i\neq j}^{k}\left(gX_{[a_i,b_i,n]s_1}-gX_{[a_j,b_j,n]s_2}\right)^2$$

Then we have:

$$
\begin{aligned}
E(A) =& \frac{1}{2M(M-1)n^2} \sum_{s_1 \neq s_2}^{M} \sum_{i=1}^{k} E\left[\left(gX_{[a_i,b_i,n]s_1} - gX_{[a_i,b_i,n]s_2}\right)^2\right] \\
=& \frac{1}{2M(M-1)n^2} \sum_{s_1 \neq s_2}^{M} \sum_{i=1}^{k} E\left[(g^2)(X^2_{[a_i,b_i,n]s_1} - 2X_{[a_i,b_i,n]s_1}X_{[a_i,b_i,n]s_2} + X^2_{[a_i,b_i,n]s_2})\right] \\
=& \frac{1}{2n^2} \sum_{i=1}^{k} (g^2)(2E(X^2_{[a_i,b_i,n]}) - 2E(X_{[a_i,b_i,n]})E(X_{[a_i,b_i,n]})) \\
=& \frac{1}{n^2} \sum_{i=1}^{k} (g^2)\left[E(X^2_{[a_i,b_i,n]}) - (E(X_{[a_i,b_i,n]}))^2\right] \\
=& \left(\frac{g}{n}\right)^2 \left[\sum_{i=1}^{n} \frac{1}{g}(E(X^2_{(i)}) - \sum_{i=1}^{k} (E(X_{[a_i,b_i,n]}))^2\right] \\
=& \left(\frac{g}{n^2}\right) nE\left(X^2\right) - \left(\frac{g}{n}\right)^2 \sum_{i=1}^{k} (E(X_{[a_i,b_i,n]}))^2 \\
=& \left(\frac{g}{n^2}\right) n(\sigma^2 + \mu^2) - \left(\frac{g}{n}\right)^2 \sum_{i=1}^{k} (E(X_{[a_i,b_i,n]}))^2 \\
=& \left(\frac{g}{n}\right) (\sigma^2 + \mu^2) - \left(\frac{g}{n}\right)^2 \sum_{i=1}^{k} (E(X_{[a_i,b_i,n]}))^2.
\end{aligned}
$$

Also,

$$
\begin{aligned}
E(B) =& \frac{1}{2M(M-1)n^2} \sum_{s_1 \neq s_2}^{M} \sum_{i \neq j}^{k} E\left[\left(gX_{[a_i,b_i,n]s_1} - gX_{[a_j,b_j,n]s_2}\right)^2\right] \\
=& \frac{1}{2M(M-1)n^2} \sum_{s_1 \neq s_2}^{M} \sum_{i \neq j}^{k} (g^2) E(X^2_{[a_i,b_i,n]s_1} - 2X_{[a_j,b_j,n]s_1}X_{[a_j,b_j,n]s_2} + X^2_{[a_j,b_j,n]s_2}) \\
=& \frac{1}{2M(M-1)n^2} M(M-1)(g^2) \left(2(k-1)\sum_{i=1}^{k} E(X^2_{[a_i,b_i,n]})\right. \\
& \left. -2\sum_{i \neq j}^{k} E(X_{[a_i,b_i,n]})E(X_{[a_i,b_i,n]})\right) \\
=& \frac{1}{n^2}(g^2) \left((k-1)\sum_{i=1}^{k} E(X^2_{[a_i,b_i,n]}) - \sum_{i,j=1}^{k} E(X_{[a_i,b_i,n]})E(X_{[a_i,b_i,n]})\right. \\
& \left. + \sum_{i=1}^{k} (E(X_{[a_i,b_i,n]}))^2\right)
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{n^2}(g^2)\left((k-1)\frac{n}{g}(\sigma^2+\mu^2)-\left(\sum_{i=1}^{k}E(X_{[a_i,b_i,n]})\right)^2+\sum_{i=1}^{k}(E(X_{[a_i,b_i,n]}))^2\right)\\
=&\frac{1}{n^2}(g^2)\left((k-1)\frac{n}{g}(\sigma^2+\mu^2)-\left(\frac{1}{g}\sum_{i=1}^{n}E(X_{(i)})\right)^2+\sum_{i=1}^{k}(E(X_{[a_i,b_i,n]}))^2\right)\\
=&\frac{g(k-1)}{n}(\sigma^2+\mu^2)-\frac{g^2}{n^2}\left(\frac{n}{g}\mu\right)^2+\left(\frac{g}{n}\right)^2\sum_{i=1}^{k}(E(X_{[a_i,b_i,n]}))^2\\
=&\frac{n-g}{n}(\sigma^2+\mu^2)-\mu^2+\left(\frac{g}{n}\right)^2\sum_{i=1}^{k}(E(X_{[a_i,b_i,n]}))^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E(\hat{\sigma}^2_{UKRPSS}) =&E(A+B)\\
=&\left(\frac{g}{n}\right)(\sigma^2+\mu^2)-\left(\frac{g}{n}\right)^2\sum_{i=1}^{k}(E(X_{[a_i,b_i,n]}))^2+\frac{n-g}{n}(\sigma^2+\mu^2)-\mu^2\\
&+\left(\frac{g}{n}\right)^2\sum_{i=1}^{k}(E(X_{[a_i,b_i,n]}))^2\\
=&\sigma^2+\mu^2-\mu^2\\
=&\sigma^2.
\end{aligned}
$$

□

2.4 Estimation of the Distribution Function

The natural estimator of the distribution function is the empirical distribution function given by:

$$
\hat{F}_{UKRPSS}(x)=\frac{1}{Mk}\sum_{s=1}^{M}\sum_{j=1}^{k}I(X_{[a_j,b_j,n]s}\leq x), \tag{2.4.1}
$$

where $I(\cdot)$ is the indicator function.

The following lemma gives the properties of $I(X_{[a_j,b_j,n]s}\leq x)$.

Lemma 2.4. *For a given x, we have the following properties for $I(X_{[a_j,b_j,n]} \leq x)$:*

1. $Var\left[I(X_{[a_j,b_j,n]} \leq x)\right] = \frac{1}{g}\sum_{i=a_j}^{b_j} F_{(i:n)}(x) - \frac{1}{g^2}\left(\sum_{i=a_j}^{b_j} F_{(i:n)}(x)\right)^2$

2. *For $(a_j, b_j) \neq (a_l, b_l)$*

$$\begin{aligned} Cov(I(X_{[a_j,b_j,n]} \leq x), I(X_{[a_l,b_l,n]} \leq x)) =& \frac{1}{g^2}\left[\sum_{q=a_j}^{b_j}\sum_{r=a_l}^{b_l} F_{(q,r:n)}(x,x)\right. \\ & \left. - \left(\sum_{q=a_j}^{b_j} F_{(q:n)}(x)\right)\left(\sum_{r=a_l}^{b_l} F_{(r:n)}(x)\right)\right] \end{aligned}$$

   where $F_{(q,r:n)}(x,x) = P(X_{(q:n)} \leq x, X_{(r:n)} \leq x)$.

3. $$\begin{aligned} Cov(I(X_{[a_j,b_j,n]} \leq x), I(X_{[a_j,b_j,n]} \leq y)) =& \frac{1}{g}\sum_{q=a_j}^{b_j} F_{(q:n)}(x \wedge y) \\ & - \left(\frac{1}{g}\sum_{q=a_j}^{b_j} F_{(q:n)}(x)\right)\left(\frac{1}{g}\sum_{q=a_j}^{b_j} F_{(q:n)}(y)\right) \end{aligned}$$

where $g = b_j - a_j + 1$, $F_{(q:n)}$ is the distribution function of the q-th order statistic from an SRS of size n, and $F_{(q,r:n)}$ is the joint distribution function of the q-th and r-th order statistics from an SRS of size n.

Proof.

1.

$$\begin{aligned} Var\left[I(X_{[a_j,b_j,n]} \leq x)\right] =& E\left[\left(I(X_{[a_j,b_j,n]} \leq x)\right)^2\right] - E^2\left[I(X_{[a_j,b_j,n]} \leq x)\right] \\ =& P(X_{[a_j,b_j,n]} \leq x) - \left[Pr(X_{[a_j,b_j,n]} \leq x)\right]^2 \\ =& F_{[a_j,b_j,n]}(x) - F^2_{[a_j,b_j,n]}(x) \\ =& \frac{1}{g}\sum_{i=a_j}^{b_j} F_{(i:n)}(x) - \frac{1}{g^2}\left(\sum_{i=a_j}^{b_j} F_{(i:n)}(x)\right)^2. \end{aligned}$$

The last step results from Lemma 2.2.

2. For $(a_j, b_j) \neq (a_l, b_l)$,

$$Cov(I(X_{[a_j,b_j,n]} \leq x), I(X_{[a_l,b_l,n]} \leq x))$$

$$\begin{aligned}
=&E\left[I(X_{[a_j,b_j,n]} \leq x)I(X_{[a_l,b_l,n]} \leq x)\right] - E\left[I(X_{[a_j,b_j,n]} \leq x)\right] E\left[I(X_{[a_l,b_l,n]} \leq x)\right] \\
=&E\left[E\left[I(X_{(q:n)} \leq x)I(X_{(r:n)} \leq x)|X_{[a_j,b_j,n]} \sim X_{(q:n)}, X_{[a_l,b_l,n]} \sim X_{(r:n)}\right]\right] \\
&- F_{[a_j,b_j,n]}(x)F_{[a_l,b_l,n]}(x) \\
=&\frac{1}{g^2}\sum_{q=a_j}^{b_j}\sum_{r=a_l}^{b_l} F_{(q,r:n)}(x,x) - \frac{1}{g^2}\left(\sum_{q=a_j}^{b_j} F_{(q:n)}(x)\right)\left(\sum_{r=a_l}^{b_l} F_{(r:n)}(x)\right) \\
=&\frac{1}{g^2}\left[\sum_{q=a_j}^{b_j}\sum_{r=a_l}^{b_l} F_{(q\vee r:n)}(x) - \left(\sum_{q=a_j}^{b_j} F_{(q:n)}(x)\right)\left(\sum_{r=a_l}^{b_l} F_{(r:n)}(x)\right)\right]
\end{aligned}$$

3. $Cov(I(X_{[a_j,b_j,n]} \leq x), I(X_{[a_l,b_l,n]} \leq y)) =$

$$\begin{aligned}
&E\left[I(X_{[a_j,b_j,n]} \leq x)I(X_{[a_j,b_j,n]} \leq y)\right] - E\left[I(X_{[a_j,b_j,n]} \leq x)\right] E\left[I(X_{[a_j,b_j,n]} \leq y)\right] \\
=&E\left[E\left[I(X_{(q:n)} \leq x \wedge y)|X_{[a_j,b_j,n]} = X_{(q:n)}\right]\right] - F_{[a_j,b_j,n]}(x)F_{[a_j,b_j,n]}(y) \\
=&\frac{1}{g}\sum_{q=a_j}^{b_j} F_{(q:n)}(x \wedge y) - \left(\frac{1}{g}\sum_{q=a_j}^{b_j} F_{(q:n)}(x)\right)\left(\frac{1}{g}\sum_{q=a_j}^{b_j} F_{(q:n)}(y)\right)
\end{aligned}$$

□

Theorem 2.4. *Suppose an UKPRSS(k, g, M) is drawn from a distribution F. Then, for a given x:*

1. *$\hat{F}_{UKRPSS}(x)$ is an unbiased estimator for $F(x)$.*

2. *The variance of $\hat{F}_{UKRPSS}(x)$ is:*

$$Var(\hat{F}_{UKRPSS}(x)) = Var(\hat{F}_{SRS}(x)) + \frac{2}{Nk}\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x) - \frac{k-1}{N}F^2(x)$$

where $\hat{F}_{SRS}(x)$ is the empirical distribution function based on an SRS of size $N = kM$.

3. $Var(\hat{F}_{UKRPSS}(x)) \leq Var(\hat{F}_{SRS}(x))$.

Proof.

1. Unbiasedness

$$\begin{aligned}
E\left[\hat{F}_{UKRPSS}(x)\right] =& E\left[\frac{1}{Mk}\sum_{s=1}^{M}\sum_{j=1}^{k} I(X_{[a_j,b_j,n]s} \leq x)\right] \\
=& \frac{1}{Mk}\sum_{s=1}^{M}\sum_{j=1}^{k} E\left[I(X_{[a_j,b_j,n]s} \leq x)\right] \\
=& \frac{1}{Mk}\sum_{s=1}^{M}\sum_{j=1}^{k} F_{[a_j,b_j,n]s}(x) \\
=& \frac{1}{Mk}\sum_{s=1}^{M}\sum_{j=1}^{k} \frac{1}{g}\sum_{i=a_j}^{b_j} F_{(i:n)}(x) \\
=& \frac{1}{Mkg}\sum_{s=1}^{M}\sum_{i=1}^{n} F_{(i:n)}(x) \\
=& \frac{1}{Mkg}\sum_{s=1}^{M} nF(x) \\
=& \frac{1}{kg} kgF(x) \\
=& F(x).
\end{aligned}$$

2. Variance

$$\begin{aligned}
Var\left(\hat{F}_{UKRPSS}(x)\right) =& Var\left(\frac{1}{Mk}\sum_{s=1}^{M}\sum_{j=1}^{k} I(X_{[a_j,b_j,n]s} \leq x)\right) \\
=& \frac{1}{(Mk)^2}\sum_{s=1}^{M} Var\left[\sum_{j=1}^{k} I(X_{[a_j,b_j,n]s} \leq x)\right] \\
=& \frac{1}{Mk^2} Var\left[\sum_{j=1}^{k} I(X_{[a_j,b_j,n]} \leq x)\right]
\end{aligned}$$

Case: $k = 1$

Then $N = M$ and $g = n$. Thus,

$$\begin{aligned} Var\left(\hat{F}_{UKRPSS}(x)\right) =& \frac{1}{N} Var\left[I\left(X_{[1,n,n]} \leq x\right)\right] \\ =& \frac{1}{N}\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} F_{(i:n)}(x) - \frac{1}{n^2}\left(\sum_{i=1}^{n} F_{(i:n)}(x)\right)^2\right] \\ =& \frac{1}{N}\left(F(x) - F^2(x)\right) \\ =& Var(\hat{F}_{SRS}(x)) \end{aligned}$$

Case: $k = n$

Then $N = Mn, g = 1$ and $X_{[a_i,b_i,n]} = X_{(i:n)}$

$$\begin{aligned} Var\left(\hat{F}_{UKRPSS}(x)\right) =& \frac{1}{Mn^2}\left[\sum_{i=1}^{n}\sum_{j=i}^{n} Cov\left(I(X_{[a_i,b_i,n]} \leq x), I(X_{[a_j,b_j,n]} \leq x)\right)\right] \\ =& \frac{1}{Mn^2}\left[\sum_{i=1}^{n}\sum_{j=i}^{n} Cov\left(I(X_{(i:n)} \leq x), I(X_{(j:n)} \leq x)\right)\right] \\ =& \frac{1}{Mn^2}\left[\sum_{i=1}^{n}\sum_{j=i}^{n} Cov\left(I(X_i \leq x), I(X_j \leq x)\right)\right] \\ =& \frac{1}{Mn^2} nVar\left[I(X \leq x)\right] \\ =& \frac{1}{Mn}\left(F(x) - F^2(x)\right) \\ =& Var(\hat{F}_{SRS}(x)). \end{aligned}$$

Case: $1 < k < n$

$$\begin{aligned} Var\left(\hat{F}_{UKRPSS}(x)(x)\right) =& \frac{1}{Mk^2}\left[\sum_{j=1}^{k} Var\left(I(X_{[a_j,b_j,n]} \leq x)\right)\right. \\ & \left. +2\sum_{1\leq q<r\leq k} Cov\left(I(X_{[a_q,b_q,n]} \leq x), I(X_{[a_r,b_r,n]} \leq x)\right)\right] \end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{Mk^2}\left[\sum_{j=1}^{k}\frac{1}{g}\sum_{i=a_j}^{b_j}F_{(i:n)}(x)-\sum_{j=1}^{k}F^2_{[a_j,b_j,n]}(x)\right.\\
&\left.+2\sum_{1\le q<r\le k}\left(F_{[a_r,b_r,n]}(x)-F_{[a_r,b_r,n]}(x)F_{[a_q,b_q,n]}(x)\right)\right]\\
=&\frac{1}{Mk^2}\left[\sum_{j=1}^{k}\frac{1}{g}\sum_{i=a_j}^{b_j}F_{(i:n)}(x)-\sum_{j=1}^{k}F^2_{[a_j,b_j,n]}(x)\right.\\
&\left.+2\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)-2\sum_{1\le q<r\le k}F_{[a_r,b_r,n]}(x)F_{[a_q,b_q,n]}(x)\right]\\
=&\frac{1}{Mk^2}\left[kF(x)-\sum_{j=1}^{k}F^2_{[a_j,b_j,n]}(x)+2\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)\right.\\
&\left.-\left(\left(\frac{1}{g^2}\right)\sum_{1\le i,j\le n}F_{(i:n)}(x)F_{(j:n)}(x)-\sum_{j=1}^{k}F^2_{[a_j,b_j,n]}(x)\right)\right]\\
=&\frac{1}{Mk^2}\left[kF(x)+2\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)\right.\\
&\left.-\left(\frac{1}{g^2}\right)\sum_{1\le i,j\le n}F_{(i:n)}(x)F_{(j:n)}(x)\right]\\
=&\frac{1}{Mk^2}\left[kF(x)+2\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)-k^2F^2(x)\right]\\
=&\frac{1}{Mk^2}\left[kF(x)-kF^2(x)+2\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)\right.\\
&\left.-(k^2-k)F^2(x)\right]\\
=&Var(\hat{F}_{SRS}(x))+\frac{2}{Mk^2}\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)-\frac{k^2-k}{Mk^2}F^2(x)\\
=&Var(\hat{F}_{SRS}(x))+\frac{2}{Nk}\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x)-\frac{k-1}{N}F^2(x)
\end{aligned}
$$

3. The difference between $Var(\hat{F}_{UKPRSS}(x))$ and $Var(\hat{F}_{SRS}(x))$, at a given x, is

$$\Delta(x) = \frac{2}{Nk}\sum_{r=2}^{k}(r-1)F_{[a_r,b_r,n]}(x) - \frac{k-1}{N}F^2(x).$$

When $\Delta(x) < 0$, then $Var(\hat{F}_{UKPRSS}(x)) < Var(\hat{F}_{SRS}(x))$

We can rewrite $\Delta(x)$ as

$$\begin{aligned}\Delta(x) =& \frac{2}{Nk}\sum_{r=2}^{k}(r-1)\frac{1}{g}\sum_{i=a_r}^{b_r}F_{(i:n)}(x) - \frac{k-1}{N}F^2(x)\\ =& \frac{2}{Nn}\sum_{r=2}^{k}(r-1)\sum_{i=a_r}^{b_r}F_{(i:n)}(x) - \frac{k-1}{N}F^2(x)\end{aligned}$$

We now consider:

$$\frac{2}{n}\sum_{r=2}^{k}(r-1)\sum_{i=a_r}^{b_r}F_{(i:n)}(x) - (k-1)F^2(x)$$

and denote for fixed n and x,

$$\begin{aligned}w(k) =& \frac{2}{n}\sum_{r=2}^{k}(r-1)\sum_{i=a_r}^{b_r}F_{(i:n)}(x),\\ h(k) =& (k-1)F^2(x).\end{aligned}$$

It is clear that $w(1) = h(1) = 0$. Also, from the proof of $Var\left(\hat{F}_{UKPRSS}(x)\right)$, we have that when $k = n$, $Var\left(\hat{F}_{UKPRSS}(x)\right) = Var\left(\hat{F}_{SRS}(x)\right)$. This implies that $w(n) = h(n)$. An alternative proof is provided at the end of this chapter.

Now note that $(k-1)F^2(x)$ is linear in k and increases at a rate of $F^2(x)$. Also, notice that $w(k)$ is monotonically increasing in k. As k increases, the summation includes more terms. In addition, if $F_{(i:n)}$ was already included in the summation, its coefficient either stays the same, or increases. See Table 2.3 for examples.

So the differences between consecutive values of k is increasing. Along with the fact that $w(1) = h(1)$, there is a unique $k > 1$ such that $w(k') \geq h(k')$. Because $w(n) = h(n)$, it must be that for $1 < k < n$, $w(k) < h(k)$. □

k	$r-1$	terms
2	1	$F_{(\frac{n}{2}+1:n)}(x), \ldots, F_{(n:n)}(x)$
3	1	$F_{(\frac{n}{3}+1:n)}(x), \ldots, F_{(\frac{2n}{3}:n)}(x)$
	2	$F_{(\frac{2n}{3}+1:n)}(x), \ldots, F_{(n:n)}(x)$
4	1	$F_{(\frac{n}{4}+1:n)}(x), \ldots, F_{(\frac{2n}{4}:n)}(x)$
	2	$F_{(\frac{2n}{4}+1:n)}(x), \ldots, F_{(\frac{3n}{4}:n)}(x)$
	3	$F_{(\frac{3n}{4}+1:n)}(x), \ldots, F_{(n:n)}(x)$

Table 2.3: Example values of $(r-1)$ and corresponding distributions of order statistics for different values of k for $\Delta(x)$

.

Before moving on, we first present an asymptotic two sided confidence interval for $F(x)$.

Corollary 2.2. *As $M \to \infty$, an asymptotic $100(1-\alpha)\%$ two sided confidence interval at the point x is given by:*

$$\hat{F}_{UKPRSS}(x) \pm z^* \sqrt{\widehat{Var}(\hat{F}_{UKPRSS}(x))}$$

where z^* is the critical value for $(1-\frac{\alpha}{2})$ from the standard normal distribution and

$$\begin{aligned}\widehat{Var}(\hat{F}_{UKPRSS}(x)) =& \frac{1}{Mk}\left(\hat{F}_{UKPRSS}(x) - \hat{F}^2{}_{UKPRSS}(x)\right) - \frac{k-1}{Mk}\hat{F}^2{}_{UKPRSS}(x) \\ &+ \frac{2}{Mk^2}\sum_{r=2}^{k}\left[\frac{(r-1)}{g}\sum_{i=a_r}^{b_r}\sum_{j=i}^{n}\binom{n}{j}\hat{F}^j{}_{UKPRSS}(x)\left(1-\hat{F}_{UKPRSS}(x)\right)^{n-j}\right]\end{aligned}$$

(i.e. $\widehat{Var}(\hat{F}_{UKPRSS}(x))$ is $Var(\hat{F}_{UKPRSS}(x))$ with $F(x)$ replaced by $\hat{F}_{UKPRSS}(x)$.)

Proof. Let x be fixed. Then, by the central limit theorem,

$$\left(\hat{F}_{UKPRSS}(x) - F(x)\right) \xrightarrow{d} \mathcal{N}(0, Var(\hat{F}_{UKPRSS}(x))$$

Also note that as $M \to \infty$, $Var(\hat{F}_{UKPRSS}(x)) \to 0$. Thus, $\hat{F}_{UKPRSS}(x) \xrightarrow{P} F(x)$, where $\xrightarrow{P}$ denotes convergence in probability, and $\widehat{Var}(\hat{F}_{UKPRSS}(x)) \to Var(\hat{F}_{UKPRSS}(x))$. By

Slutsky's theorem,

$$\frac{\left(\hat{F}_{UKPRSS}(x) - F(x)\right)}{\sqrt{\widehat{Var}(\hat{F}_{UKPRSS}(x))}} \xrightarrow{d} \mathcal{N}(0,1)$$

and the confidence interval follows. □

2.4.1 Simulation Study: Standard Normal

Table 2.5 shows both the calculated variance and the simulated counterparts of $\hat{F}_{UKRPSS}(x)$ when x is the first, second, and third quartiles of the standard normal distribution. Simulated variance was based on 10,000 repetitions. Total sample size was 12 ($N = 12$) for all situations.

Table 2.6 shows the ratio of the simulated variances to the variances of $\hat{F}_{SRS}(x)$ (shown in Table 2.4). As expected, when group size, g, is n or 1, the variance is equal to that of $\hat{F}_{SRS}(x)$, and is lower when $1 < g < n$.

	$x =$ Q1	$x =$ Median	$x =$ Q3
$Var(\hat{F}_{SRS}(x))$	0.0156	0.02083	0.0156

Table 2.4: Variances of $Var(\hat{F}_{SRS}(x))$.

	$x =$ Q1		$x =$ Median		$x =$ Q3	
Group Size	Simulated Variance	Calculated Variance	Simulated Variance	Calculated Variance	Simulated Variance	Calculated Variance
12	0.0155	0.0156	0.0208	0.0208	0.0158	0.0156
6	0.0105	0.0107	0.0093	0.0094	0.0106	0.0107
4	0.0082	0.0084	0.0093	0.0096	0.0082	0.0084
3	0.0082	0.0083	0.0102	0.0100	0.0085	0.0083
2	0.0094	0.0095	0.0122	0.0122	0.0096	0.0095
1	0.0161	0.0156	0.0207	0.0208	0.0155	0.0156

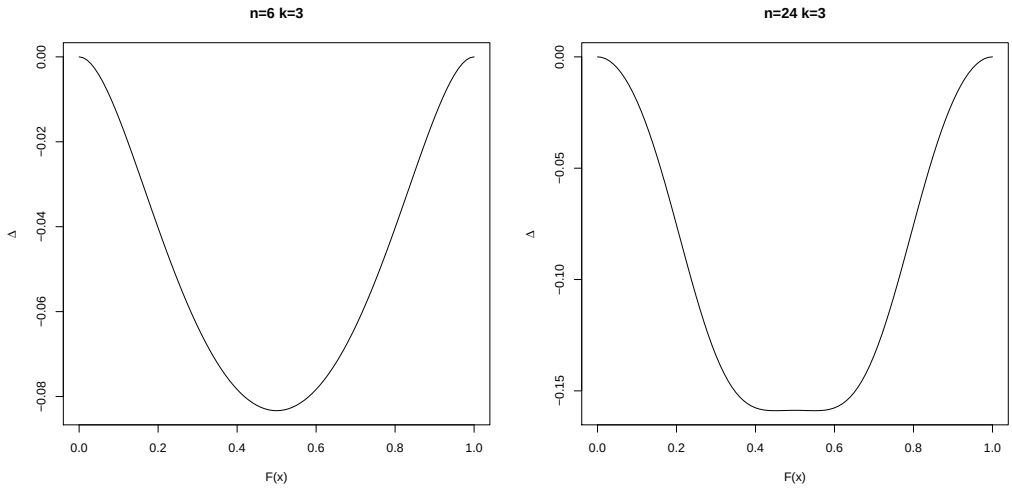
Table 2.5: Theoretical and Simulated Variances for $Var(\hat{F}_{UKRPSS}(x))$. Simulation consisted of 10000 repititions.

Group Size	x = Q1	x = Median	x = Q3
12	1.0000	1.0000	1.0000
6	1.4659	2.2165	1.4659
4	1.8597	2.1695	1.8597
3	1.8805	2.0771	1.8805
2	1.6364	1.7143	1.6364
1	1.0000	1.0000	1.0000

Table 2.6: $Var(\hat{F}_{SRS}(x))/Var(\hat{F}_{UKRPSS}(x))$ using theoretical variances.

2.4.2 Calculations of Reduction in Variance

For $6 \leq n \leq 100$, n not prime, and for all corresponding values of k such that $k \neq 1$ or n, $\Delta(x)$ was calculated at each percentile (i.e. $F(x) = 0, 0.1, 0.2, ..., 1$). For all cases checked, $\Delta(x) \leq 0$, as expected. Figure 2.4.2 presents four graphs indicative of the characteristic plots seen.



(a) (b)

(c) (d)

Figure 2.1: Example plots of Δ for some combinations of n and k.

2.5 Supplemental Proof

Lemma 2.5. $\frac{2}{n}\sum_{r=2}^{n}(r-1)F_{(r:n)}(x) = (n-1)F^2(x)$

Proof. Consider

$$\sum_{\substack{1\le i,j\le n\\ i\neq j}} F_{(i,j:n)}(x,x).$$

Because all pairs (i,j) with $i \neq j$ are considered, we have that:

$$\begin{aligned}\sum_{\substack{1\le i,j\le n\\ i\neq j}} F_{(i,j:n)}(x,x) =& \sum_{\substack{1\le i,j\le n\\ i\neq j}} F_{i,j}(x,x)\\ =& \sum_{\substack{1\le i,j\le n\\ i\neq j}} F_i(x)F_j(x)\\ =& n(n-1)F^2(x). \end{aligned} \tag{2.5.1}$$

Now, for $i < j$, we have $F_{(i,j:n)}(x,x) = F_{(j:n)}(x)$ (David and Nagaraja, 2003). This gives:

$$\sum_{\substack{1\le i,j\le n\\ i\neq j}} F_{(i,j:n)}(x) = 2\sum_{r=2}^{n}(r-1)F_{(r:n)} \tag{2.5.2}$$

Setting (2.5.1) equal to (2.5.2) gives the result. □

CHAPTER 3:

BALANCED K-TUPLE PARTIALLY RANK-ORDERED SET SAMPLING

In this section, we generalize the Uniform KPRSS sampling plan to Balanced KPRSS. As before, we derive estimators of the population mean, variance, and distribution function under the assumption of perfect ranking of groups for each partially rank-ordered set.

3.1 Sampling Procedure

For a Balanced k-Tuple Partially Ranked-Ordered Set Sample of size N, using a set size of n, the sampling procedure is as follows:

1. Randomly select n units.

2. Divide the n units into G groups, each of size g, and order the groups. This forms a partially ranked-ordered set of size n. (Note: We now have $n = Gg$, whereas in the last chapter, $n = kg$.)

3. Select k of the groups.

4. From each of the selected k groups, randomly select one unit for measurement.

5. Repeat steps $1-4$ with a new selection of n units and a different set of k groups. Cycling through all the $\binom{G}{k}$ possible choices of k groups. This completes one cycle.

6. Repeat for M cycles, until $N\ (= kM\binom{G}{k})$ units have been measured.

This will be denoted as BKPRSS(M, G, g, k). The number of units that need to be screened for a balanced KPRSS is:

$$U = nM\binom{G}{k} = \frac{Nn}{k}$$

When $g = 1$ and $k = 1$, then $G = n$. Thus, BKPRSS$(M, n, 1, 1)$ is a balanced ranked set sample with set size n and M replicates. The number of units needing to be screened is:

$$U = nM\binom{n}{1} = Mn^2$$

which is the number of units needing to be screened for a balanced ranked set sample with set size n and M cycles.

Additionally, because each partially rank-ordered set are independent, but groups within a partially rank-ordered set are not, RSS has no covariance terms from partially rank-ordered groups. In fact, any Balanced KPRSS sample with $k = 1$ has no covariance between partially rank-ordered groups. On the other end, Uniform KPRSS is a Balanced KPRSS with $k = G$, and includes that highest amount of covariance between partially ranked-ordered groups.

We will use the following notation:

1. $(1, 2, \ldots, k : s_m)$: is the k-tuple corresponding to the s-th simple random sample in the m-th cycle.

2. $[a_j, b_j, n]s_m$: is the partially ordered group that contains the a_j-th through b_j-th order statistics from the s-th simple random sample of size n in the m-th cycle. Note, for

$m \neq n$, $[a_j, b_j, n]s_m$ need not be the same partially ordered group as $[a_j, b_j, n]s_n$. It may be that s_m and s_n have different associated k-tuples.

3. $X_{[a_j,b_j,n]s}$: is a draw from $[a_j, b_j, n]s$.

3.2 Estimation of the Population Mean

Theorem 3.1. *Suppose a Balanced K-Tuple Partially Ordered Set Sample is drawn from a distribution with mean μ. Let*

$$\overline{X}_{BKPRSS} = \frac{1}{Mk\binom{G}{k}} \sum_{m=1}^{M} \sum_{s_m=1}^{\binom{G}{k}} \sum_{j=1}^{k} X_{[a_j,b_j,n]s_m}$$

Then:

1. $E\left(\overline{X}_{BKPRSS}\right) = \mu$

2. $$Var\left(\overline{X}_{BKPRSS}\right) = \left(\frac{k-1}{g(G-1)}\right)\sigma^2_{\overline{X}_{SRS}} + \left(\frac{1}{N}\right)E\left(X^2\right) - \left(\frac{1}{g^2GN}\right)\sum_{i=1}^{G}\left(\sum_{a_i}^{b_i}\mu_{(i:n)}\right)^2 - \left(\frac{(k-1)}{g^2G(G-1)N}\right)\sum_{i=1}^{G}\sum_{q,r=a_i}^{b_i}\sigma_{(q,r:n)}$$

   where $\sigma_{(q,r:n)} = Cov\left(X_{(q:n)}, X_{(r:n)}\right)$ and $\sigma^2_{\overline{X}_{SRS}}$ is the variance of an SRS of size N.

3. $Var(\overline{X}_{BKPRSS}) \leq Var(\overline{X}_{SRS})$

Proof.

1. Expectation

$$\begin{aligned}
E\left(\overline{X}_{BKPRSS}\right) =& E\left(\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\sum_{j=1}^{k}X_{[a_j,b_j,n]s_m}\right)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\sum_{j=1}^{k}E\left(X_{[a_j,b_j,n]s_m=1}\right)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m}^{\binom{G}{k}}\sum_{j=1}^{k}\left(\frac{1}{g}\sum_{i=a_j}^{b_j}\mu_{(i:n)}\right)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\left(\frac{1}{g}\binom{G-1}{k-1}\sum_{i=1}^{n}\mu_{(i:n)}\right)\\
=&\frac{1}{Mk\binom{G}{k}}\left(\frac{Mn}{g}\binom{G-1}{k-1}\mu\right)\\
=&\frac{k!(G-k)!}{kG!}\left(\frac{n}{g}\right)\left(\frac{(G-1)!}{(k-1)!(G-k)!}\right)\mu\\
=&\mu
\end{aligned}$$

2. Variance

$$\begin{aligned}
Var\left(\overline{X}_{BKPRSS}\right) =& Var\left(\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\sum_{j=1}^{k}X_{[a_j,b_j,n]s_m}\right)\\
=&\left(\frac{1}{Mk\binom{G}{k}}\right)^2\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}Var\left(\sum_{j=1}^{k}X_{[a_j,b_j,n]s_m}\right)\\
=&\left(\frac{1}{Mk\binom{G}{k}}\right)^2\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\left[\sum_{j=1}^{k}Var\left(X_{[a_j,b_j,n]s_m}\right)\right.\\
&\left.+\sum_{\substack{1\le i,j\le k\\ i\neq j}}Cov\left(X_{[a_i,b_i,n]s_m},X_{[a_j,b_j,n]s_m}\right)\right] \qquad (3.2.1)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{1}{Mk\binom{G}{k}}\right)^2 \sum_{m=1}^{M} \left[\sum_{s_m=1}^{\binom{G}{k}} \sum_{j=1}^{k} Var\left(X_{[a_j,b_j,n]s_m}\right) \right. \\
&\quad \left. + \sum_{s_m=1}^{\binom{G}{k}} \sum_{\substack{1\le i,j\le k \\ i\neq j}} Cov\left(X_{[a_i,b_i,n]s_m}, X_{[a_j,b_j,n]s_m}\right) \right] \\
&= \left(\frac{1}{Mk\binom{G}{k}}\right)^2 \sum_{m=1}^{M} \left[\binom{G-1}{k-1} \sum_{j=1}^{G} Var\left(X_{[a_j,b_j,n]s_m}\right) \right. \\
&\quad \left. + \binom{G-2}{k-2} \sum_{\substack{1\le i,j\le G \\ i\neq j}} Cov\left(X_{[a_i,b_i,n]s_m}, X_{[a_j,b_j,n]s_m}\right) \right] \\
&= \left(\frac{1}{Mk\binom{G}{k}}\right)^2 \sum_{m=1}^{M} \left[\binom{G-1}{k-1} \sum_{j=1}^{G} \left\{ \frac{1}{g} \sum_{i=a_j}^{b_j} E\left(X_{(i)}^2\right) - \frac{1}{g^2} \left(\sum_{i=a_j}^{b_j} \mu_{(i:n)} \right)^2 \right\} \right. \\
&\quad \left. + \binom{G-2}{k-2} \frac{1}{g^2} \left\{ \sum_{1\le i,j\le n} \sigma_{(i,j:n)} - \sum_{i=1}^{G} \sum_{a_i\le q,r\le b_i} \sigma_{(q,r:n)} \right\} \right] \\
&= \left(\frac{1}{Mk^2\binom{G}{k}^2}\right) \left[\binom{G-1}{k-1} \frac{n}{g} E\left(X^2\right) - \binom{G-1}{k-1} \sum_{j=1}^{G} \frac{1}{g^2} \left(\sum_{i=a_j}^{b_j} \mu_{(i:n)} \right)^2 \right. \\
&\quad \left. + \binom{G-2}{k-2} \frac{n}{g^2} \sigma^2 - \binom{G-2}{k-2} \frac{1}{g^2} \sum_{i=1}^{G} \sum_{a_i\le q,r\le b_i} \sigma_{(q,r:n)} \right] \\
&= \left(\frac{k-1}{g(G-1)}\right) \sigma^2_{\overline{X}_{SRS}} - \left(\frac{1}{g^2 GMk\binom{G}{k}}\right) \sum_{j=1}^{G} \left(\sum_{i=a_j}^{b_j} \mu_{(i:n)} \right)^2 \\
&\quad + \left(\frac{1}{Mk\binom{G}{k}}\right) E\left(X^2\right) - \left(\frac{(k-1)}{g^2 G(G-1)Mk\binom{G}{k}}\right) \sum_{i=1}^{G} \sum_{a_i\le q,r\le b_i} \sigma_{(q,r:n)} \\
&= \left(\frac{k-1}{g(G-1)}\right) \sigma^2_{\overline{X}_{SRS}} + \left(\frac{1}{N}\right) E\left(X^2\right) - \left(\frac{1}{g^2 GN}\right) \sum_{j=1}^{G} \left(\sum_{i=a_j}^{b_j} \mu_{(i:n)} \right)^2 \\
&\quad - \left(\frac{(k-1)}{g^2 G(G-1)N}\right) \sum_{i=1}^{G} \sum_{a_i\le q,r\le b_i} \sigma_{(q,r:n)}
\end{aligned}
$$

3. Because the covariance of order statistics is always non-negative (Bickel, 1967), equations

2.2.2 and 3.2.1 imply that $Var(\overline{X}_{BKPRSS})$ is maximized when $k = G$. Then we have,

$$\frac{(k-1)}{g^2G(G-1)N} = \frac{1}{g^2GN} = \frac{1}{g^2kN} = \frac{k}{g^2k^2N} = \frac{k}{n^2N}$$

and $\sigma^2_{\overline{X}_{BKPRSS}} = \sigma^2_{\overline{X}_{UKPRSS}}$. The result then follows from Theorem 2.1 □

3.2.1 Simulation Studies

Simulation from Standard Normal

Balanced KPRSS samples of size 60 (N=60) were drawn from the standard normal distribution using a partially rank-ordered set of size of 12 (n=12) and various combinations k-Tuple sizes and group sizes. Ranking was assumed to be perfect. Table 3.1 below shows the simulated and theoretical variance of $\overline{X}_{BKPRSS}$ and the corresponding number of units screened. Simulated variance was calculated from 10,000 balanced KPRSS samples of size 60. The column for $\mu_{X_{BKPRSS}}$ represents the average of all 10,000 sample means.

Columns 6 and 7 of Table 3.1 show close agreement between the simulated variance of $\overline{X}_{BKPRSS}$ and the theoretical variance. As expected, for a fixed number of groups, the smaller the value of k, the lower the variance, but also the larger the number of units screened. Also, because the number of units screened is $\frac{Nn}{k}$, the greatest reduction in screening pool size is when k is largest. When $k = n$, Var($\overline{X}_{BKPRSS}$) =Var($\overline{X}_{SRS}$). This can be seen in the second row of Table 3.1.

All other rows in the table show a reduction in variance when compared to $\overline{X}_{SRS}$. Overall, the first row has the greatest reduction in variance. This row is a balanced ranked set sample of size 60. The second row shows equal variance between $\overline{X}_{BKPRSS}$ and $\overline{X}_{SRS}$. This is because no ranking information is used when all n units of a ranked set are measured. Thus, when $g = 1$ and $k = n$, one has essentially collected a simple random sample.

Number of Groups (G)	Group Size (g)	Tuple Size (k)	Number of Screened Units	$\mu_{X_{BKPRSS}}$	Simulated Var($\overline{X}_{BKPRSS}$)	Theoretical Var($\overline{X}_{BKPRSS}$)	Var($\overline{X}_{SRS}$)
12	1	1	720	0.0006	0.0030	**0.0030**	0.0167
12	1	12	60	0.0003	0.0164	0.0167	0.0167
6	2	1	720	0.0003	0.0036	0.0035	0.0167
6	2	2	360	-0.0008	0.0046	0.0047	0.0167
6	2	3	240	0.0002	0.0057	0.0058	0.0167
6	2	4	180	-0.0010	0.0071	0.0070	0.0167
6	2	5	144	-0.0002	0.0080	0.0081	0.0167
6	2	6	120	0.0001	0.0093	0.0093	0.0167
4	3	1	720	0.0000	0.0043	0.0042	0.0167
4	3	2	360	-0.0004	0.0054	0.0057	0.0167
4	3	3	240	0.0010	0.0063	0.0072	0.0167
4	3	4	180	-0.0002	0.0075	0.0075	0.0167
3	4	1	720	-0.0002	0.0051	0.0051	0.0167
3	4	2	360	-0.0008	0.0062	0.0069	0.0167
3	4	3	240	0.0002	0.0071	0.0072	0.0167
2	6	1	720	0.0002	0.0074	0.0073	0.0167
2	6	2	360	-0.0006	0.0084	0.0082	0.0167

Table 3.1: Number of units screened, theoretical variance calculation and simulated variance for $\overline{X}_{BKPRSS}$ from 10,000 samples from the standard normal distribution. $\mu_{X_{BKPRSS}}$ is the average of all 10,000 sample means. Assumes perfect ranking.

Bivariate Normal

To examine the effect of ranking error arising from using a correlated ranking variable, we ran another simulation study. We assumed that the ranking variable and the variable of interest came from a bivariate normal with

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Samples of size $N = 60$ were generated using partially rank-ordered sets of size $n = 12$. Correlations used were $\rho = 0, 0.4, 0.9, 1,$ and -1. Table 3.2 gives the variance of $\overline{X}_{BKPRSS}$ based on 10,000 simulations. It can be seen that as correlation increases, variance of the sample mean decreases. When $\rho = -1$, the variance is the same as $\rho = 1$, even though every element in a partially rank-ordered set would be mis-ranked. This is because in a balanced sampling plan, each partially rank-ordered group is equally represented.

		Correlation				
Group Size (g)	k	0	0.4	0.9	1	-1
1	1	0.0168	0.0148	0.0056	0.0030	0.0030
1	12	0.0168	0.0161	0.0164	0.0164	0.0168
2	1	0.0165	0.0144	0.0061	0.0036	0.0036
2	2	0.0169	0.0148	0.0070	0.0046	0.0046
2	3	0.0167	0.0147	0.0079	0.0058	0.0060
2	4	0.0168	0.0152	0.0087	0.0069	0.0069
2	5	0.0166	0.0160	0.0097	0.0078	0.0082
2	6	0.0170	0.0152	0.0107	0.0090	0.0093
3	1	0.0169	0.0142	0.0067	0.0043	0.0042
3	2	0.0162	0.0146	0.0075	0.0054	0.0052
3	3	0.0166	0.0153	0.0083	0.0065	0.0063
3	4	0.0165	0.0152	0.0091	0.0077	0.0074
4	1	0.0164	0.0142	0.0076	0.0052	0.0052
4	2	0.0166	0.0149	0.0081	0.0063	0.0062
4	3	0.0162	0.0152	0.0091	0.0074	0.0070
6	1	0.0166	0.0152	0.0093	0.0072	0.0072
6	2	0.0172	0.0153	0.0099	0.0083	0.0081

Table 3.2: N=60. n=12. $Var(\overline{X}_{BKPRSS})$ for different correlations between ranking variable and variable of interest from multivariate normal. $Var(\overline{X}_{SRS}) = 0.0168$.

3.2.2 Data Example

The following analysis was conducted on a dataset originally from Platt et al. (1988) and reproduced by Chen et al. (2004). The original dataset contained data on seven variables from 399 conifer trees (*Pinus palustris*). The reproduction contained only height and diameter (in centimeters) at breast height of 396 trees. For convenience, the data is reproduced in Appendix A.

Treating the 396 trees as the population, summary statistics can be found in Table 3.3. The variable of interest is tree height. To calculate the simulated variances 10,000 Balanced KPRSS samples were collected. Table 3.4 used perfect judgement ranking and Table 3.5 used diameter at chest height for ranking. For each sample, the sample mean was calculated. The column for $\mu_{X_{BKPRSS}}$ represents the average of all 10,000 sample means and Var($\overline{X}_{BKPRSS}$) is the variance of all 10,000 sample averages.

	Diameter	Height
Mean	21.03586	52.6768
Variance	309.8617	3253.44
Correlation	0.90451	

Table 3.3: Summary Statistics for Tree Data

The results mirror those from the standard normal simulation. The first rows of Tables 3.4 and 3.5 represents a balanced ranked sample and had the greatest reduction in variance. Almost all cases displayed an increase in $\text{Var}(\overline{X}_{BKPRSS})$ when using diameter as the ranking variable. This is because correlation between height and diameter is not 1 and ranking errors are introduced. The one case without an increase in variance was row two, which corresponds to a simple random sample. This, again, is as expected.

Comparison of percent increase of variance from perfect ranking vs. when ranking error is present are shown in Table 3.6. It can be seen from this table that for a fixed k, as group size increases, percent increase in variance decreases. Logically, ranking errors are most likely to occur between units within a group as opposed to between groups. The larger the group size, the lower the number of ranking decisions that must be made and the lower the effect on variance of estimators. The largest percent increase was for RSS (i.e. $k = 1$, $g = 1$), which is consistent with this reasoning.

Table 3.6 also shows the average ranking error from each simulation. For every partially rank-ordered set, the percent of mis-ranked units was calculated. Even though the correlation between diameter and height is 0.90451, the average ranking error was above 51%. Even still, $\text{Var}(\overline{X}_{BKPRSS})$ was still lower than $\text{Var}(\overline{X}_{SRS})$ in all cases.

The natural question is why not always take an RSS sample if it will still be the best performer? RSS still requires a large number of screening units. Also, because partially

rank-ordered set sampling is less affected by ranking error its behavior is more predictable when percentage of ranking error is unknown.

Number of Groups (G)	Group Size (g)	Tuple Size (k)	Number of Screened Units	$\mu_{X_{BKPRSS}}$	Simulated Var($\overline{X}_{BKPRSS}$)	Var($\overline{X}_{SRS}$)
12	1	1	720	52.6906	**12.4928**	54.3577
12	1	12	60	52.6717	**52.3682**	53.7814
6	2	1	720	52.6239	19.7745	53.4859
6	2	2	360	52.6476	14.9862	55.2967
6	2	3	240	52.6617	21.374	53.5750
6	2	4	180	52.7359	23.9309	54.4332
6	2	5	144	52.6300	27.1521	53.6479
6	2	6	120	52.8168	30.3349	53.7193
4	3	1	720	52.7177	18.4819	54.5369
4	3	2	360	52.7161	21.384	53.9773
4	3	3	240	52.6445	23.6163	54.7887
4	3	4	180	52.5226	26.5035	55.4425
3	4	1	720	52.6747	22.4046	54.2788
3	4	2	360	52.7008	25.3948	55.1779
3	4	3	240	52.6672	26.8973	54.3005
2	6	1	720	52.7451	32.1422	54.7417
2	6	2	360	52.6106	33.5909	54.9225

Table 3.4: N = 60. n = 12. Simulated variance for $\overline{X}_{BKPRSS}$ calculated from 10,000 samples. Samples based on perfect ranking. Theoretical Var($\overline{X}_{SRS}$)=54.2240.

Number of Groups (G)	Group Size (g)	Tuple Size (k)	Number of Screened Units	$\mu_{X_{BKPRSS}}$	Simulated Var($\overline{X}_{BKPRSS}$)	Var($\overline{X}_{SRS}$)
12	1	1	720	52.6568	**17.8595**	54.3577
12	1	12	60	52.6688	**52.3682**	53.7814
6	2	1	720	52.5791	19.4969	53.4859
6	2	2	360	52.6761	22.6101	55.2967
6	2	3	240	52.6489	25.638	53.5750
6	2	4	180	52.6762	27.4404	54.4332
6	2	5	144	52.6218	30.4031	53.6479
6	2	6	120	52.8306	33.6704	53.7193
4	3	1	720	52.7469	21.8289	54.5369
4	3	2	360	52.7044	24.9219	53.9773
4	3	3	240	52.6668	27.1102	54.7887
4	3	4	180	52.5098	29.4588	55.4425
3	4	1	720	52.6643	25.3052	54.2788
3	4	2	360	52.6991	27.5771	55.1780
3	4	3	240	52.7161	28.6992	54.3005
2	6	1	720	52.7505	33.5367	54.7417
2	6	2	360	52.5911	35.15	54.9225

Table 3.5: N = 60. n = 12. Diameter at chest height used as the ranking variable. Simulated variance for $\overline{X}_{BKPRSS}$ calculated from 10,000 samples. Theoretical Var($\overline{X}_{SRS}$)=54.2240.

Number of Groups (G)	Group Size (g)	k-Tuple Size (k)	Percent Increase In Variance	Average Ranking Error Percentage
12	1	1	**0.4296**	0.5168
12	1	12	**0**	0.5176
6	2	1	0.3010	0.5170
6	2	2	0.2428	0.5174
6	2	3	0.1995	0.5164
6	2	4	0.1467	0.5169
6	2	5	0.1197	0.5165
6	2	6	0.1100	0.5169
4	3	1	0.1811	0.5164
4	3	2	0.1654	0.5179
4	3	3	0.1479	0.5176
4	3	4	0.1115	0.5166
3	4	1	0.1295	0.5169
3	4	2	0.0859	0.5165
3	4	3	0.0670	0.5171
2	6	1	0.0434	0.5166
2	6	2	0.0464	0.5166

Table 3.6: N = 60. n = 12. Percent Increase in Variance Using Diameter as Ranking Variable VS Perfect Ranking.

3.3 Estimation of the Population Variance

As in the last chapter, an estimator of the population variance will be constructed through the weighted sum of within-group variance and between-group variance estimators. Let

$$A = \frac{1}{4M\binom{G-1}{k-1}\left[\binom{G-1}{k-1}-1\right]n^2}\sum_{m=1}^{M}\sum_{r_m\neq s_m}^{\binom{G}{k}}\sum_{i=1}^{G}\left(gX_{[a_i,b_i,n]r_m}-gX_{[a_i,b_i,n]s_m}\right)^2$$

$$B = \frac{1}{4M(M-1)\binom{G-1}{k-1}^2n^2}\sum_{m_1\neq m_2}^{M}\sum_{r_{m_1},s_{m_2}}^{\binom{G}{k}}\sum_{i=1}^{G}\left(gX_{[a_i,b_i,n]r_{m_1}}-gX_{[a_i,b_i,n]s_{m_2}}\right)^2$$

$$C = \frac{1}{4M\left[\binom{G-1}{k-1}^2-\binom{G-2}{k-2}\right]n^2}\sum_{m=1}^{M}\sum_{r_m\neq s_m}^{\binom{G}{k}}\sum_{i\neq j}^{G}\left(gX_{[a_i,b_i,n]r_m}-gX_{[a_j,b_j,n]s_m}\right)^2$$

$$D = \frac{1}{4M(M-1)\binom{G-1}{k-1}^2n^2}\sum_{m_1\neq m_2}^{M}\sum_{r_{m_1},s_{m_2}}^{\binom{G}{k}}\sum_{i\neq j}^{G}\left(gX_{[a_i,b_i,n]r_{m_1}}-gX_{[a_j,b_j,n]s_{m_1}}\right)^2$$

Terms A and B estimate the within-group variance within a cycle and the within-group variance between cycles, respectively. Terms C and D estimate the between-group variance within a cycle and the between-group variance between cycles, respectively. The estimator of the population variance is then:

$$\hat{\sigma}^2_{BKPRSS} = A + B + C + D$$

Theorem 3.2.

$\hat{\sigma}^2_{BKPRSS}$ is an unbiased estimator for the population variance.

Proof.

$$E(A) = \frac{1}{4M\binom{G-1}{k-1}\left[\binom{G-1}{k-1}-1\right]n^2}E\left[\sum_{m=1}^{M}\sum_{r_m\neq s_m}^{\binom{G}{k}}\sum_{i=1}^{G}\left(gX_{[a_i,b_i,n]r_m}-gX_{[a_i,b_i,n]s_m}\right)^2\right]$$

$$
\begin{aligned}
=&\frac{g^2 M\binom{G-1}{k-1}\left[\binom{G-1}{k-1}-1\right]}{4M\binom{G-1}{k-1}\left[\binom{G-1}{k-1}-1\right]n^2}\sum_{i=1}^{G}E\left(X^2_{[a_i,b_i,n]}\right)-2E^2\left(X_{[a_i,b_i,n]}\right)+E\left(X^2_{[a_i,b_i,n]}\right)\\
=&\frac{2g^2}{4n^2}\sum_{i=1}^{G}\left[E\left(X^2_{[a_i,b_i,n]}\right)-E^2\left(X^2_{[a_i,b_i,n]}\right)\right]\\
=&\frac{2g^2}{4n^2}\left[\sum_{i=1}^{n}\frac{1}{g}E\left(X^2_{(i)}\right)-\sum_{i=1}^{G}E^2\left(X^2_{[a_i,b_i,n]}\right)\right]\\
=&\frac{g}{2n^2}nE\left(X^2\right)-\frac{g^2}{2n^2}\sum_{i=1}^{G}E^2\left(X^2_{[a_i,b_i,n]}\right)\\
=&\frac{g}{2n}E\left(X^2\right)-\frac{g^2}{2n^2}\sum_{i=1}^{G}E^2\left(X^2_{[a_i,b_i,n]}\right).
\end{aligned}
$$

The second equality is because in any particular cycle and any particular group $X_{[a_i,b_i,n]}$, the number of k-tuples the group will appear in is $\binom{G-1}{k-1}$. So the number of times the expectation of this group will appear is $\mathrm{M}\binom{G-1}{k-1}\left[\binom{G-1}{k-1}-1\right]$.

$$
\begin{aligned}
E(B)=&\frac{1}{4M(M-1)\binom{G-1}{k-1}^2n^2}E\left[\sum_{m_1\neq m_2}^{M}\sum_{r_{m_1},s_{m_2}}^{\binom{G}{k}}\sum_{i=1}^{G}\left(gX_{[a_i,b_i,n]r_{m_1}}-gX_{[a_i,b_i,n]s_{m_2}}\right)^2\right]\\
=&\frac{g^2M(M-1)\binom{G-1}{k-1}^2}{4M(M-1)\binom{G-1}{k-1}^2n^2}\sum_{i=1}^{G}E\left(X^2_{[a_i,b_i,n]}\right)-2E^2\left(X_{[a_i,b_i,n]}\right)+E\left(X^2_{[a_i,b_i,n]}\right)\\
=&\frac{2g^2}{4n^2}\sum_{i=1}^{G}\left[E\left(X^2_{[a_i,b_i,n]}\right)-E^2\left(X_{[a_i,b_i,n]}\right)\right]\\
=&\frac{2g^2}{4n^2}\left[\sum_{i=1}^{n}\frac{1}{g}E\left(X^2_{(i)}\right)-\sum_{i=1}^{G}E^2\left(X_{[a_i,b_i,n]}\right)\right]\\
=&\frac{g}{2n^2}nE\left(X^2\right)-\frac{g^2}{2n^2}\sum_{i=1}^{G}E^2\left(X_{[a_i,b_i,n]}\right)\\
=&\frac{g}{2n}E\left(X^2\right)-\frac{g^2}{2n^2}\sum_{i=1}^{G}E^2\left(X_{[a_i,b_i,n]}\right)\\
=&E(A).
\end{aligned}
$$

$$
\begin{aligned}
E(C) =& \frac{1}{4M\left[\binom{G-1}{k-1}^2 - \binom{G-2}{k-2}\right]n^2} E\left[\sum_{m=1}^{M}\sum_{r_m\neq s_m}^{\binom{G}{k}}\sum_{i\neq j}^{G}\left(gX_{[a_i,b_i,n]r_m} - gX_{[a_j,b_j,n]s_m}\right)^2\right] \\
=& \frac{g^2 M\left[\binom{G-1}{k-1}^2 - \binom{G-2}{k-2}\right]}{4M\left[\binom{G-1}{k-1}^2 - \binom{G-2}{k-2}\right]n^2}\left[\sum_{i=1}^{G} 2(G-1)E\left(X^2_{[a_i,b_i,n]}\right)\right. \\
& \left. -2\sum_{i\neq j}^{G} E\left(X_{[a_i,b_i,n]}\right)E\left(X_{[a_j,b_j,n]}\right)\right] \\
=& \frac{g^2}{2n^2}\left[\sum_{i=1}^{G}(G-1)E\left(X^2_{[a_i,b_i,n]}\right) - \sum_{i\neq j}^{G} E\left(X_{[a_i,b_i,n]}\right)E\left(X_{[a_j,b_j,n]}\right)\right] \\
=& \frac{g^2(G-1)}{2n^2}\sum_{i=1}^{G} E\left(X^2_{[a_i,b_i,n]}\right) - \frac{g^2}{2n^2}\left[\sum_{i,j}^{G} E\left(X_{[a_i,b_i,n]}\right)E\left(X_{[a_j,b_j,n]}\right) - \sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right)\right] \\
=& \frac{g(G-1)}{2n^2} nE\left(X^2\right) - \frac{g^2}{2n^2}\left[\frac{1}{g^2}\sum_{i,j}^{n} E\left(X_{(i)}\right)E\left(X_{(j)}\right) - \sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right)\right] \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{1}{2n^2}\left[\sum_{i}^{n} E\left(X_{(i)}\right)\right]^2 + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{1}{2n^2}\left[\sum_{i}^{n} E\left(X\right)\right]^2 + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{1}{2n^2}\left[n\mu\right]^2 + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{\mu^2}{2} + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right).
\end{aligned}
$$

The reasoning for the coefficient in the second line is this: for any pair of partially ordered groups, the number of k-tuples in which each group will appear is $\binom{G-1}{k-1}$. So there will be $\binom{G-1}{k-1}^2$ possible replicates for the difference of these two groups. But, differences within a k-tuple are not considered and the number of k tuples containing both groups is $\binom{G-2}{k-2}$ must be subtracted off. This happens $(G-1)$ times for any particular group since it is compared to all remaining groups.

$$
\begin{aligned}
E(D) =& \frac{1}{4M(M-1)\binom{G-1}{k-1}^2 n^2} E\left[\sum_{m_1 \neq m_2}^{M} \sum_{r_{m_1}, s_{m_2}}^{\binom{G}{k}} \sum_{i \neq j}^{G} \left(g X_{[a_i,b_i,n] r_{m_1}} - g X_{[a_j,b_j,n] s_{m_1}}\right)^2\right] \\
=& \frac{g^2 M(M-1)\binom{G-1}{k-1}^2}{4M(M-1)\binom{G-1}{k-1}^2 n^2} \left[\sum_{i=1}^{G} 2(G-1) E\left(X^2_{[a_i,b_i,n]}\right) - 2\sum_{i \neq j}^{G} E\left(X_{[a_i,b_i,n]}\right) E\left(X_{[a_j,b_j,n]}\right)\right] \\
=& \frac{g^2}{2n^2}\left[\sum_{i=1}^{G} (G-1) E\left(X^2_{[a_i,b_i,n]}\right) - \sum_{i \neq j}^{G} E\left(X_{[a_i,b_i,n]}\right) E\left(X_{[a_j,b_j,n]}\right)\right] \\
=& \frac{g^2(G-1)}{2n^2} \sum_{i=1}^{G} E\left(X^2_{[a_i,b_i,n]}\right) - \frac{g^2}{2n^2}\left[\sum_{i,j}^{G} E\left(X_{[a_i,b_i,n]}\right) E\left(X_{[a_j,b_j,n]}\right)\right. \\
& \left. - \sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right)\right] \\
=& \frac{g(G-1)}{2n^2} n E\left(X^2\right) - \frac{g^2}{2n^2}\left[\frac{1}{g^2}\sum_{i,j}^{n} E\left(X_{(i)}\right) E\left(X_{(j)}\right) - \sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right)\right] \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{1}{2n^2}\left[\sum_{i}^{n} E\left(X_{(i)}\right)\right]^2 + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{1}{2n^2}\left[\sum_{i}^{n} E\left(X\right)\right]^2 + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{1}{2n^2}\left[n\mu\right]^2 + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& \frac{g(G-1)}{2n} E\left(X^2\right) - \frac{\mu^2}{2} + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) \\
=& E(C).
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
E\left(\hat{\sigma}^2_{BKPRSS}\right) =& E\left(A\right) + E\left(B\right) + E\left(C\right) + E\left(D\right) \\
=& 2[E(A) + E(C)] \\
=& 2\left[\frac{g}{2n} E\left(X^2\right) - \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right) + \frac{g(G-1)}{2n} E\left(X^2\right)\right. \\
& \left. - \frac{\mu^2}{2} + \frac{g^2}{2n^2}\sum_{i=1}^{G} E^2\left(X_{[a_i,b_i,n]}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= 2\left[\frac{gG}{2n}E\left(X^2\right) - \frac{\mu^2}{2}\right] \\
&= E\left(X^2\right) - \mu^2 \\
&= \sigma^2.
\end{aligned}
$$

□

3.4 Estimation of the Distribution Function

We now turn to estimation of the distribution function. The natural estimator is the empirical distribution function,

$$\hat{F}_{BKPRSS}(x) = \frac{1}{Mk\binom{G}{k}} \sum_{m=1}^{M} \sum_{s_m=1}^{\binom{G}{k}} \sum_{j=1}^{k} I(X_{[a_j,b_j,n]s_m} \leq x)$$

where $I(\cdot)$ is the indicator function.

Theorem 3.3. *Suppose a Balanced k-Tuple Partially Rank-Ordered Set Sample is drawn from a distribution F. Then, for a given x:*

1. *$\hat{F}_{BKPRSS}(x)$ is an unbiased estimator of $F(x)$.*

2. *The variance of $\hat{F}_{BKPRSS}(x)$ is:*

$$
\begin{aligned}
Var\left(\hat{F}_{BKPRSS}(x)\right) =& Var(\hat{F}_{SRS}(x)) + \frac{\binom{G-2}{k-2} - \binom{G-1}{k-1}}{Mk^2\binom{G}{k}^2} \sum_{j=1}^{G} \frac{1}{g^2} \left(\sum_{i=a_j}^{b_j} F_{(i:n)}(x)\right)^2 \\
&+ \frac{2\binom{G-2}{k-2}}{Mk^2\binom{G}{k}^2} \sum_{j=2}^{G} (j-1) F_{[a_j,b_j,n]}(x) - \frac{\binom{G-2}{k-2}G^2 - \binom{G-1}{k-1}G}{Mk^2\binom{G}{k}^2} F^2(x).
\end{aligned}
$$

3. *$Var(\hat{F}_{BKPRSS}(x)) \leq Var(\hat{F}_{SRS}(x))$.*

Proof.

1. Expectation

$$\begin{aligned}
E\left(\hat{F}_{BKPRSS}(x)\right) =& E\left(\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\sum_{j=1}^{k} I(X_{[a_j,b_j,n]s_m}\leq x)\right)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\sum_{j=1}^{k} E\left(I(X_{[a_j,b_j,n]s_m}\leq x)\right)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\binom{G-1}{k-1}\sum_{j=1}^{G} F_{[a_j,b_j,n]}(x)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\binom{G-1}{k-1}\sum_{j=1}^{G}\sum_{i=a_j}^{b_j}\frac{1}{g}F_{(i:n)}(x)\\
=&\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\binom{G-1}{k-1}\frac{n}{g}F(x)\\
=&\frac{M(k!)(G-k)!}{Mk(G!)}\left(\frac{(G-1)!}{(k-1)!(G-k)!}\right)GF(x)\\
=&F(x)
\end{aligned}$$

2. Variance

$$\begin{aligned}
Var\left(\hat{F}(x)\right) =& Var\left(\frac{1}{Mk\binom{G}{k}}\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\sum_{j=1}^{k} I(X_{[a_j,b_j,n]s_m}\leq x)\right)\\
=&\left(\frac{1}{Mk\binom{G}{k}}\right)^2\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}} Var\left(\sum_{j=1}^{k} I(X_{[a_j,b_j,n]s_m}\leq x)\right)\\
=&\left(\frac{1}{Mk\binom{G}{k}}\right)^2\sum_{m=1}^{M}\sum_{s_m=1}^{\binom{G}{k}}\left[\sum_{j=1}^{k} Var\left(I(X_{[a_j,b_j,n]s_m}\leq x)\right)\right.\\
&\left.+\sum_{\substack{1\leq i,j\leq k\\ i\neq j}} Cov\left(I(X_{[a_i,b_i,n]s_m}\leq x), I(X_{[a_j,b_j,n]s_m}\leq x)\right)\right] \qquad (3.4.1)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{1}{Mk^2\binom{G}{k}^2}\right)\left[\sum_{s=1}^{\binom{G}{k}}\sum_{j=1}^{k} Var\left(I(X_{[a_j,b_j,n]s}\le x)\right)\right.\\
&\quad \left.+\sum_{s=1}^{\binom{G}{k}}\sum_{\substack{1\le i,j\le k\\ i\neq j}} Cov\left(I(X_{[a_i,b_i,n]s}\le x), I(X_{[a_j,b_j,n]s}\le x)\right)\right]\\
&= \left(\frac{1}{Mk^2\binom{G}{k}^2}\right)\left[\binom{G-1}{k-1}\sum_{j=1}^{G} Var\left(I(X_{[a_j,b_j,n]}\le x)\right)\right.\\
&\quad \left.+\binom{G-2}{k-2}\sum_{\substack{1\le i,j\le G\\ i\neq j}} Cov\left(I(X_{[a_i,b_i,n]}\le x), I(X_{[a_j,b_j,n]}\le x)\right)\right]\\
&= \frac{1}{Mk^2\binom{G}{k}^2}\left[\binom{G-1}{k-1}\left(\frac{1}{g}\sum_{j=1}^{G}\sum_{i=a_j}^{b_j} F_{(i:n)}(x) - \sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j} F_{(i:n)}(x)\right)^2\right)\right.\\
&\quad +\binom{G-2}{k-2}\left\{2\sum_{1\le i<j\le G} F_{[a_i,b_i,n],[a_j,b_j,n]}(x,x)\right.\\
&\quad \left.\left.-\sum_{\substack{1\le i,j\le G\\ i\neq j}} F_{[a_i,b_i,n]}(x)F_{[a_j,b_j,n]}(x)\right\}\right]\\
&= \frac{1}{Mk^2\binom{G}{k}^2}\left[\binom{G-1}{k-1}\left(\frac{1}{g}\sum_{j=1}^{G}\sum_{i=a_j}^{b_j} F_{(i:n)}(x) - \sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j} F_{(i:n)}(x)\right)^2\right)\right.\\
&\quad +\binom{G-2}{k-2}\left\{2\sum_{j=2}^{G}(j-1)F_{[a_j,b_j,n]}(x)\right.\\
&\quad \left.\left.-\left(\sum_{1\le i,j\le n}\frac{1}{g^2}F_{(i:n)}(x)F_{(j:n)}(x) - \sum_{j=1}^{G}F^2_{[a_j,b_j,n]}\right)\right\}\right]\\
&= \frac{1}{Mk^2\binom{G}{k}^2}\left[\binom{G-1}{k-1}\frac{n}{g}F(x) - \binom{G-1}{k-1}\sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j} F_{(i:n)}(x)\right)^2\right.\\
&\quad +2\binom{G-2}{k-2}\sum_{j=2}^{G}(j-1)F_{[a_j,b_j,n]}(x)\\
&\quad \left.-\binom{G-2}{k-2}\frac{n^2}{g^2}F^2(x) + \binom{G-2}{k-2}\sum_{j=1}^{G}F^2_{[a_j,b_j,n]}(x)\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{Mk^2\binom{G}{k}^2}\left[G\binom{G-1}{k-1}F(x)\right.\\
&+\left\{\binom{G-2}{k-2}-\binom{G-1}{k-1}\right\}\sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j}F_{(i:n)}(x)\right)^2\\
&\left.+2\binom{G-2}{k-2}\sum_{j=2}^{G}(j-1)F_{[a_j,b_j,n]}(x)-\binom{G-2}{k-2}\frac{n^2}{g^2}F^2(x)\right]\\
=&\frac{F(x)-F^2(x)}{Mk\binom{G}{k}}+\frac{\binom{G-2}{k-2}-\binom{G-1}{k-1}}{Mk^2\binom{G}{k}^2}\sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j}F_{(i:n)}(x)\right)^2\\
&+\frac{2\binom{G-2}{k-2}}{Mk^2\binom{G}{k}^2}\sum_{j=2}^{G}(j-1)F_{[a_j,b_j,n]}(x)-\frac{\binom{G-2}{k-2}G^2-\binom{G-1}{k-1}G}{Mk^2\binom{G}{k}^2}F^2(x)\\
=&Var(\hat{F}_{SRS}(x))+\frac{\binom{G-2}{k-2}-\binom{G-1}{k-1}}{Mk^2\binom{G}{k}^2}\sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j}F_{(i:n)}(x)\right)^2\\
&+\frac{2\binom{G-2}{k-2}}{Mk^2\binom{G}{k}^2}\sum_{j=2}^{G}(j-1)F_{[a_j,b_j,n]}(x)-\frac{\binom{G-2}{k-2}G^2-\binom{G-1}{k-1}G}{Mk^2\binom{G}{k}^2}F^2(x).
\end{aligned}
$$

3. Because each partially rank-ordered set is independent, $Var(\hat{F}_{UKPRSS}(x))$ is maximized when the number of covariance terms in Equation 3.4.1 is maximized. This occurs when $k = G$. The result then follows from the fact that $Var(\hat{F}_{UKPRSS}(x)) \leq Var(\hat{F}_{SRS}(x))$. □

As with UKPRSS, we have an asymptotic confidence interval for $F(x)$:

Corollary 3.1. *As $M \to \infty$, an asymptotic $100(1-\alpha)\%$ two-sided confidence interval for $F(x)$ at the point x is given by:*

$$\hat{F}_{BKPRSS}(x) \pm z^* \sqrt{\widehat{Var}(\hat{F}_{BKPRSS}(x))}$$

where z^* is the critical value for $(1-\frac{\alpha}{2})$ from the standard normal distribution and

$$
\begin{aligned}
&\widehat{Var}(\hat{F}_{BKPRSS}(x)) = \\
&\frac{\hat{F}_{BKPRSS}(x) - \hat{F}^2_{BKPRSS}(x)}{Mk\binom{G}{k}} - \frac{\binom{G-2}{k-2}G^2 - \binom{G-1}{k-1}G}{Mk^2\binom{G}{k}^2}\hat{F}^2_{BKPRSS}(x) \\
&+ \frac{\binom{G-2}{k-2} - \binom{G-1}{k-1}}{Mk^2\binom{G}{k}^2}\sum_{j=1}^{G}\frac{1}{g^2}\left(\sum_{i=a_j}^{b_j}\sum_{r=i}^{n}\binom{n}{r}\hat{F}^r{}_{BKPRSS}(x)\left(1-\hat{F}_{BKPRSS}(x)\right)^{n-r}\right)^2 \\
&+ \frac{2\binom{G-2}{k-2}}{Mk^2\binom{G}{k}^2}\sum_{j=2}^{G}(j-1)\left(\frac{1}{g}\sum_{i=a_j}^{b_j}\sum_{r=i}^{n}\binom{n}{r}\hat{F}^r{}_{BKPRSS}(x)\left(1-\hat{F}_{BKPRSS}(x)\right)^{n-r}\right)
\end{aligned}
$$

Proof. Notice that as $M \to \infty$, $\text{Var}(\hat{F}_{BKPRSS}(x)) \to 0$. Using arguments similar to the proof of Corollary 2.2, the confidence interval follows from the central limit theorem and Slutsky's theorem. □

CHAPTER 4:

GENERAL K-TUPLE PARTIALLY RANK-ORDERED SET SAMPLING

In this chapter, we generalize k-Tuple Partially Rank-Ordered Set Sampling. Unlike the balanced case in Chapter 3, where all partially rank-ordered set sizes are the same and all k-tuples get equal representation, we allow everything to vary in this case. Even the group sizes may vary. An estimator of the distribution function is proposed and its asymptotic properties studied.

4.1 Sampling Method

The general KPRSS (GKPRSS) plan for a sample of size N is as follows:

1. For the s-th SRS, randomly draw n_s units. The value of n_s can differ for different values of s.

2. Form the s-th partially rank-ordered set by dividing the n_s units into G_s partially rank-ordered groups. Each group may vary in size.

3. Select k_s of the G_s groups. From each of the k_s groups randomly, select one unit for measurement.

4. Repeat T times until N units have been measured.

Notice that in this general version, the the number of groups as well as the size of each group can differ. Also, each simple random sample can vary in size. We now introduce the following notation:

1. $[a_j, b_j, n_s]s$: is the partially ordered group from the s-th SRS. The size of the SRS is n_s. The partially ordered group contains the a_j-th through b_j-th order statistics, where $1 \leq j \leq G_s$.

2. $X_{[a_j,b_j,n_s]s}$: is a draw from $[a_j, b_j, n_s]s$.

In general, an estimator of the population mean based on the sample average will not be unbiased. This is because unbiasedness depends on having equal probability of observing each order statistic. Fortunately, an asymptotically unbiased estimator of the distribution function does exist.

4.2 Estimation of the Distribution Function

Let us define the following:

1. $L \equiv$ the number of unique k-tuples.

2. $(1_l, 2_l, ..., k_l) \equiv l^{th}$ k-tuple.

3. $n_l \equiv$ the SRS size the l^{th} k-tuple was collected from.

4. $r_l \equiv$ the number of replicates of the l^{th} k-tuple.

5. $R = \sum_{l=1}^{L} r_l \equiv$ the overall number of k-tuples collected.

6. Assume that $min(r_1, ..., r_L) \to \infty$, and as a result, $\frac{r_l}{R} \to p_l$.

Now, let

$$\begin{aligned} F_l(x) =& \frac{1}{k_l} \sum_{i=1_l}^{k_l} F_{[a_i,b_i,n_l]}(x) \\ \hat{F}_l(x) =& \frac{1}{k_l} \sum_{i=1_l}^{k_l} \frac{1}{r_l} \sum_{s_l=1}^{r_l} I\left(X_{[a_i,b_i,n_l]s_l} \leq x\right). \end{aligned}$$

Note that:

$$F_{(j:n_l)}(x) = B\left(j, n_l - j + 1, F(x)\right),$$

where $B\left(j, n_l - j + 1, x\right)$ is the CDF of a beta distribution with parameters j and $n_l - j + 1$. Thus, defining

$$h_l(y) = \frac{1}{k_l} \sum_{i=1_l}^{k_l} \sum_{j=a_i}^{b_i} \frac{1}{b_i - a_i + 1} B\left(j, n_l - j + 1, y\right),$$

we can write:

$$F_l(x) = h_l \circ F(x)$$

where "$\circ$" is the composition operator. Define

$$F_h(x) = \sum_{l=1}^{L} p_l F_l(x) = h \circ F(x),$$

where

$$h(y) = \sum_{l=1}^{L} p_l h_l(y).$$

Note that $h : [0, 1] \to [0, 1]$ is continuous and strictly increasing, as it is a weighted sum of beta CDFs with positive weights. Also, $h(0) = 0$ and $h(1) = 1$. Therefore, h^{-1} exists and is unique. Hence, we have

$$F(x) = h^{-1} \circ F_h(x).$$

We can define an estimator of $F(x)$ as

$$\hat{F}(x) = h^{-1} \circ \hat{F}_h(x)$$

where

$$\hat{F}_h(x) = \frac{1}{R} \sum_{l=1}^{L} \sum_{i=1_l}^{k_l} \frac{1}{k_l} \left(\sum_{s=1}^{r_l} I\left(X_{[a_i,b_i,n_l]s} \leq x \right) \right)$$

is the empirical CDF from the GKPRSS sample.

Theorem 4.1. *Assume that as* $\min(r_1, ..., r_L) \to \infty$, $\frac{r_l}{R} \to p_l$. *Then,*

$$\sqrt{R}\left(\hat{F}(x) - F(x) \right) \xrightarrow{d} \frac{\mathbb{W}}{h' \circ F(x)}$$

where $\mathbb{W}$ *is a Gaussian process with mean 0 and covariance kernel*

$$\sum_{l=1}^{L} \frac{p_l}{k_l^2} \sum_{i=1_l}^{k_l} \sum_{j=1_l}^{k_l} K_{([a_i,b_i,n_l],[a_j,b_j,n_l])}(s,t)$$

Here,

$$K_{([a,b,n],[c,d,n])}(s,t) = \begin{cases} \frac{1}{g_1} \sum_{q=a}^{b} F_{(q:n)}(s \wedge t) - \left(\frac{1}{g_1} \sum_{q=a}^{b} F_{(q:n)}(s) \right) \left(\frac{1}{g_1} \sum_{q=a}^{b} F_{(q:n)}(t) \right) & (a,b) = (c,d) \\ \frac{1}{(g_1)(g_2)} \sum_{q=a}^{b} \sum_{r=c}^{d} F_{(q,r:n)}(s,t) - \left(\frac{1}{g_1} \sum_{q=a}^{b} F_{q:n)}(s) \right) \left(\frac{1}{g_2} \sum_{r=c}^{d} F_{(r:n)}(t) \right) & (a,b) \neq (c,d) \end{cases}$$

where $g_1 = b - a + 1$ *and* $g_2 = d - c + 1$.

Proof. Assume that as $\min(r_1, ..., r_L) \to \infty$ then $\frac{r_l}{R} \to p_l$. Then, from the theory of empirical processes, we have

$$\hat{W}_l(x) = \sqrt{r_l} \begin{pmatrix} \frac{1}{r_l} \sum_{s=1}^{r_l} I\left(X_{[a_{1_l},b_{1_l},n_l]s} \leq x \right) - F_{[a_{1_l},b_{1_l},n_l]}(x) \\ \frac{1}{r_l} \sum_{s=1}^{r_l} I\left(X_{[a_{2_l},b_{2_l},n_l]s} \leq x \right) - F_{[a_{2_l},b_{2_l},n_l]}(x) \\ \vdots \\ \frac{1}{r_l} \sum_{s=1}^{r_l} I\left(X_{[a_{k_l},b_{k_l},n_l]s} \leq x \right) - F_{[a_{k_l},b_{k_l},n_l]}(x) \end{pmatrix} \xrightarrow{\mathrm{d}} W_l(x)$$

where $I(\cdot)$ is the indicator function and $W_l(x)$ is a k_l-dimensional Gaussian process with mean $\vec{0}$ and covariance kernel

$$K_{([a,b,n],[c,d,n])}(s,t) = \begin{cases} \frac{1}{g_1}\sum_{q=a}^{b} F_{(q:n)}(s \wedge t) - \left(\frac{1}{g_1}\sum_{q=a}^{b} F_{(q:n)}(s)\right)\left(\frac{1}{g_1}\sum_{q=a}^{b} F_{(q:n)}(t)\right) & (a,b) = (c,d) \\ \frac{1}{(g_1)(g_2)}\sum_{q=a}^{b}\sum_{r=c}^{d} F_{(q,r:n)}(s,t) - \left(\frac{1}{g_1}\sum_{q=a}^{b} F_{q:n)}(s)\right)\left(\frac{1}{g_2}\sum_{r=c}^{d} F_{(r:n)}(t)\right) & (a,b) \neq (c,d) \end{cases}$$

where $g_1 = b - a + 1$ and $g_2 = b_j - a_j + 1$.

Thus,

$$\sqrt{r_l}\left(\hat{F}_l(x) - F_l(x)\right) = \frac{1}{k_l}\mathbf{1}^T_{k_l \text{ x } 1}\hat{W}_l(x) \xrightarrow{d} \frac{1}{k_l}\mathbf{1}^T_{k_l \text{ x } 1}W_l(x)$$

where $\mathbf{1}_{k \text{ x } 1}$ is the k x 1 vector of all 1s.

We have that

$$\begin{aligned}
\sqrt{R}(\hat{F}_h(x) - F_h(x)) =& \sqrt{R}\left(\frac{1}{R}\sum_{l=1}^{L}\sum_{i=1_l}^{k_l}\frac{1}{k_l}\left(\sum_{s=1}^{r_l} I\left(X_{[a_i,b_i,n_l]s} \leq x\right)\right) - \sum_{l=1}^{L} p_l F_l(x)\right) \\
=& \sqrt{R}\left(\sum_{l=1}^{L}\frac{\sqrt{r_l}^2}{R}\sum_{i=1_l}^{k_l}\frac{1}{k_l}\left(\frac{1}{r_l}\sum_{s=1}^{r_l} I\left(X_{[a_i,b_i,n_l]s} \leq x\right)\right) - \sum_{l=1}^{L}\frac{r_l}{R}F_l(x)\right) \\
=& \left(\sum_{l=1}^{L}\frac{\sqrt{r_l}}{\sqrt{R}}\sum_{i=1_l}^{k_l}\frac{\sqrt{r_l}}{k_l}\left(\frac{1}{r_l}\sum_{s=1}^{r_l} I\left(X_{[a_i,b_i,n_l]s} \leq x\right)\right)\right. \\
& \left. - \sum_{l=1}^{L}\frac{\sqrt{r_l}}{\sqrt{R}}\frac{\sqrt{r_l}}{k_l}\sum_{i=1_l}^{k_l} F_{[a_i,b_i,n_l]}(x)\right) \\
=& \sum_{l=1}^{L}\sqrt{p_l}\mathbf{1}^T_{k \text{ x } 1}\frac{1}{k_l}\hat{W}_l(x) \\
& \xrightarrow{d} \mathbb{W}
\end{aligned} \tag{4.2.1}$$

where $\mathbb{W}$, by independence of the k-tuples, is a zero mean Gaussian process with covariance kernel

$$\sum_{l=1}^{L}\frac{p_l}{k_l^2}\sum_{i=1_l}^{k_l}\sum_{j=1_l}^{k_l}\left[K_{([a_i,b_i,n_l],[a_j,b_j,n_l])}(s,t)\right]$$

Because the beta density is continuous and non-zero on $(0,1)$ the derivative of h exists and is non-zero on $(0,1)$. By the mean value theorem, we have:

$$\hat{F}_{GKPRSS}(x) - F(x) = h^{-1} \circ \hat{F}_h(x) - h^{-1} \circ F_h(x) = \frac{\hat{F}_h(x) - F_h(x)}{h' \circ \tilde{F}(x)}$$

where $\tilde{F}(x)$ is a point on the line with end points $\hat{F}(x)$ and $F(x)$.

By the Glivenko-Cantelli Theorem, we have that

$$\sup_x \left| \hat{F}_{[a,b,n_l]}(x) - F_{[a,b,n_l]}(x) \right| = \sup_x \left| \frac{1}{r_l} \sum_{s=1}^{r_l} I\left(X_{[a,b,n_l]s} \leq x\right) - F_{[a,b,n_l]}(x) \right| \xrightarrow{a.s.} 0$$

Then, as $\min(r_1, ..., r_L) \to \infty$ and $\frac{r_l}{R} \to p_l$,

$$\begin{aligned}
\sup_x \left| \hat{F}_h(x) - F_h(x) \right| &= \sup_x \left| \sum_{l=1}^{L} \frac{r_l}{R k_l} \sum_{i=1_l}^{k_l} \hat{F}_{[a,b,n_l]}(x) - \sum_{l=1}^{L} \frac{p_l}{k_l} \sum_{i=1_l}^{k_l} F_{[a,b,n_l]}(x) \right| \\
&= \sup_x \left| \sum_{l=1}^{L} \frac{p_l}{k_l} \sum_{i=1_l}^{k_l} \left(\hat{F}_{[a,b,n_l]}(x) - F_{[a,b,n_l]}(x) \right) \right. \\
&\quad \left. + \sum_{l=1}^{L} \left(\frac{r_l}{R} - p_l \right) \frac{1}{k_l} \sum_{i=1_l}^{k_l} \hat{F}_{[a,b,n_l]}(x) \right| \\
&\leq \sum_{l=1}^{L} \frac{p_l}{k_l} \sum_{i=1_l}^{k_l} \sup_x \left| \left(\hat{F}_{[a,b,n_l]}(x) - F_{[a,b,n_l]}(x) \right) \right| + 0 \\
&\xrightarrow{a.s.} 0.
\end{aligned}$$

Hence, $\hat{F}_{GKPRSS}(x) - F(x) \xrightarrow{a.s.} 0$.

Because of this, and by continuity of h' on $(0,1)$, we have:

$$\sup_x \left| h' \circ \tilde{F}(x) - h' \circ F(x) \right| \xrightarrow{a.s.} 0$$

Thus, $h' \circ \tilde{F}(x) \xrightarrow{a.s} h' \circ F(x)$. By Slutsky's Theorem and Equation 4.2.1, we have that

$$\sqrt{R} \left(\hat{F}(x) - F(x) \right) \xrightarrow{d} \frac{\mathbb{W}}{h' \circ F(x)}.$$

□

Because $\hat{F}_{GKPRSS}$ converges asymptotically to a Gaussian process, by application of the functional delta method, asymptotically unbiased estimators of any functional of the distribution function can be found. This includes the population mean and population variance.

The next section illustrates the use of an unbalanced KPRSS sample to estimate the population distribution function.

4.3 Example: Maximum and Minimum Groups

Let N and n be given. Also, let g_1 be the size of the lowest ranked group and g_2 the size of the highest ranked group. Analogous to to extreme ranked set sampling 2 (ERSS2) from Ghosh and Tiwari (2008), a member of the lowest ranked group and the highest ranked group will be randomly selected for measurement. Then,

- $L = 1$
- $[a_{1_1}, b_{1_1}, n_1]s = [1, g_1, n] \; \forall \; s$
- $[a_{2_1}, b_{2_1}, n_1]s = [n - g_2 + 1, n, n] \; \forall \; s$

Because there is only one type of k-tuple,

$$h(y) = \frac{1}{2}\left[\sum_{i=1}^{g_1} \frac{1}{g_1} B(i, n-i+1, y) + \sum_{j=n-g_2+1}^{n} \frac{1}{g_2} B(j, n-j+1, y)\right]$$

and

$$F_h(x) = \frac{1}{2}\left[\sum_{i=1}^{g_1} \frac{1}{g_1} B(i, n-i+1, F(x)) + \sum_{j=n-g_2+1}^{n} \frac{1}{g_2} B(j, n-j+1, F(x))\right].$$

Thus, $\hat{F}_{GKPRSS}(x)$ is the value of y that solves

$$h(y) - EMCDF(x) = 0$$

where $EMCDF(x)$ is the empirical cdf from the GKPRSS sample evaluated at x.

Simulations were performed with sampling from the standard normal distribution. The lowest and highest ranked groups were of the same size g. Total sample size (N) was 50 and partially ranked-ordered set size (n) was 12 in all cases. The number of replicates used was 10,000. The simulated $E\left(\hat{F}_{GKPRSS}(x)\right)$ for various values of x can be seen in Table 4.1. The table shows general agreement between $E\left(\hat{F}_{GKPRSS}(x)\right)$ and the true CDF values. The exception being $g = 1$ and $x = -0.6745$ and 0.6745. This is most likely due to using simulated data.

The simulated variances of each estimate can be found in Table 4.2. There does not seem to be a common trend amongst the columns of the table. When $x = -1.6449$ and 1.6449, variance appears to increase as group size increases. These columns also have the lowest variances. For all other columns other than $x = 0$, the variance appears to drop and then increase. For $x = 0$, the variance appears to increase and then drop and group size increases.

Because measured units were from the lowest ranked and highest ranked groups, it is logical to assume that $\hat{F}_{GKPRSS}$ would be more accurate in estimating in the tails of the distribution. This explains the behavior for $x = -1.6449$ and 1.6449 as they correspond to the 5-th and 95-th percentiles.

As group size increases, each group contains members from a larger range of the distribution. The drop and then increase for $x = \pm - 1.0364$ could be explained by this. When $g = 1$ there is no unit in the group representing the 15-th percentile. Then after a certain point, there are enough members representing other sections of the distribution that selecting that representative unit becomes less likely. Further work remains to be done to explain the behavior in the other columns of Table 4.2.

	$x =$						
g	-1.6449	-1.0364	-0.6745	0.0000	0.6745	1.0364	1.6449
1	0.0515	0.1641	0.3421	0.4999	0.6538	0.8363	0.9484
2	0.0504	0.1542	0.2728	0.4994	0.7292	0.8464	0.9497
3	0.0507	0.1525	0.2579	0.4998	0.7421	0.8480	0.9503
4	0.0504	0.1509	0.2542	0.5006	0.7468	0.8490	0.9498
6	0.0504	0.1507	0.2499	0.4994	0.7500	0.8506	0.9501
$P(X < x)$	0.05	0.15	0.25	0.50	0.75	0.85	0.95

Table 4.1: $N = 50$, $n = 12$. Simulated $E\left(\hat{F}_{GKPRSS}(x)\right)$ from 10,000 replicates from the standard normal distribution. Units selected were from the lowest and highest ranked groups only. All groups were of size g.

	$x =$						
g	-1.6449	-1.0364	-0.6745	0.0000	0.6745	1.0364	1.6449
1	0.0002	0.0040	0.0201	0.0009	0.0203	0.0037	0.0002
2	0.0003	0.0012	0.0063	0.0027	0.0062	0.0011	0.0003
3	0.0004	0.0011	0.0025	0.0039	0.0024	0.0011	0.0004
4	0.0006	0.0013	0.0020	0.0033	0.0020	0.0013	0.0006
6	0.0009	0.0021	0.0026	0.0022	0.0026	0.0021	0.0009

Table 4.2: $N = 50$, $n = 12$. Simulated $Var\left(\hat{F}_{GKPRSS}(x)\right)$ from 10,000 replicates from the standard normal distribution. Units selected were from the lowest and highest ranked groups only. All groups were of size g.

CHAPTER 5:

CONCLUSIONS AND FUTURE WORK

5.1 Conclusion

In this dissertation, we have presented k-tuple partially rank-ordered set sampling (KPRSS). This is a generalization of the partially rank-ordered set sampling scheme which allows multiple measurements from each partially rank-ordered set. Uniform, Balanced, and General KPRSS sampling schemes were presented. Under the assumption of perfect ranking between groups in each partially rank-ordered set, properties of estimates based on partially rank-ordered set amples were obtained.

For the Uniform and Balanced cases, the sample mean was shown to be unbiased with variance less than or equal to its counterpart based on an SRS of the same size. An analysis of tree data from Platt et al. (1988) provided an example of $\overline{X}_{BKPRSS}$ outperforming $\overline{X}_{SRS}$. The sample variance was shown to be biased and an unbiased estimator was presented. An unbiased estimator of the distribution function was also presented along with its variance. Like the sample mean, it was shown that $Var(\hat{F}_{BKPRSS}(x)) \leq Var(\hat{F}_{SRS}(x))$.

From the derivation of the unbiased estimates in Chapters 2 and 3, it can be seen that the unbiasedness of KPRSS estimates is based on each order statistic having equal probability of being measured. Because of this, only an asymptotically unbiased estimator of the dis-

tribution function was found for the General KPRSS plan. From this, other asymptotically unbiased estimators can be found through application of the functional delta-method.

Each variance formula presented was a functions of k and g. Screening pool size is determined by k. The smaller k is, the larger the screening pool. Group size, g, serves as a proxy for confidence of ranking ability. The larger g is in value, the lower confidence in assigning accurate ranks.

With these variables, a researcher can assess the precision of their estimates in terms of ranking confidence and screening pool size. Depending on their specific cost constraints, a researcher can then try to optimize the precision of their estimates.

5.2 Future Work

Having found the asymptotic distribution of $\hat{F}_{KPRSS}$, the first extension of this work would be to develop efficient asymptotic tests of hypotheses such as 2-sample tests.

Also, this dissertation was done under the assumption of perfect ranking of groups for each partially rank-ordered set. In practice, this assumption is only practical when collecting a General KPRSS sample. A Balanced or Uniform KPRSS sample may force tied units to be placed into different groups. Analysis of how ranking error will affect the bias and variance of estimators can be done similar to those for ranked set sampling.

Another assumption of this work is that each member of a partially ranked-ordered group is equally likely to be selected for measurement. If this is not the case, a weight for probability of selection can be added. Thus, the distribution function of a draw from a

partially rank-ordered group can then be re-written as:

$$F_{[a,b,n]}(x) = \sum_{i=a}^{b} w_i F_{(i:n)}(x)$$

where $\sum_{i=a}^{b} w_i = 1$ and $w_i \geq 0 \; \forall \; i$. A prior distribution for the w_i's may even be added to allow analysis from a Bayesian perspective.

Finally, it would be interesting to extend this idea to the case of sampling from discrete distributions, where observations are naturally tied.

APPENDIX A:

TREE DATA

Table 5.1: Tree Data Reprinted from Chen et al. (2004)

Specimen	Diameter	Height	Specimen	Diameter	Height
1	15.9	28	34	4.7	14
2	22	26	35	11	19
3	56.9	119	36	58.8	222
4	9.6	16	37	3.5	4
5	24.6	43	38	10.1	28
6	3.3	7	39	16.9	38
7	11.4	21	40	10.8	26
8	4.7	6	41	9	21
9	21.3	40	42	8	19
10	16.8	28	43	17.8	38
11	5.1	12	44	23.9	37
12	7.5	22	45	2.3	5
13	3.1	7	46	5.8	13
14	4.9	7	47	6	16
15	6.1	9	48	8.8	23
16	5.5	12	49	9.9	20
17	6.5	11	50	14.6	34
18	5.6	14	51	10.8	29
19	6.9	11	52	44.2	181
20	32.8	6	53	12.9	16
21	9.7	27	54	28	77
22	6.9	16	55	39.8	76
23	4.1	8	56	20.4	37
24	58.5	192	57	47.3	111
25	46	203	58	35.7	66
26	22.2	51	59	44.9	87
27	3.7	5	60	8.7	25
28	52.9	162	61	24.3	46
29	63.2	223	62	15.7	35
30	46.5	211	63	30.9	54
31	56.3	196	64	69.2	131
32	21.9	43	65	24.1	72
33	11	20	66	4.2	8

67	3.8	8	100	9.2	27
68	41.2	94	101	5.9	9
69	39.8	68	102	6.2	12
70	18.6	33	103	13.3	22
71	38.7	68	104	13.4	30
72	12.2	17	105	33.9	82
73	6	16	106	33.7	93
74	8	14	107	8.3	26
75	13.5	19	108	48	99
76	20.1	32	109	40.4	78
77	57.4	202	110	8.6	22
78	8.2	22	111	16	26
79	32.7	41	112	29.1	49
80	9.4	23	113	18.4	22
81	8.9	25	114	26.8	37
82	9.2	18	115	6.2	7
83	6.1	14	116	2.9	6
84	7.5	19	117	3	8
85	52.3	152	118	14.6	20
86	15.5	25	119	18.4	32
87	23.7	51	120	15	34
88	67.1	208	121	18.4	41
89	12.3	16	122	44.5	64
90	14	16	123	4.5	8
91	4.9	9	124	10.4	20
92	5.5	8	125	24	37
93	7.6	17	126	5.1	10
94	3.5	5	127	5.3	13
95	6.3	18	128	2.5	4
96	19	39	129	2.2	3
97	2.7	5	130	3.1	4
98	8.2	24	131	2.6	4
99	7.6	20	132	8.1	26

133	12.4	31	166	4.7	8
134	15.1	34	167	5.3	10
135	12.7	38	168	10.6	19
136	49	96	169	3.7	6
137	20.8	35	170	3.9	8
138	11.9	18	171	5.3	12
139	47.6	154	172	2.5	3
140	10.6	32	173	13.2	38
141	22.9	33	174	17.1	37
142	10.6	27	175	13.9	33
143	49.7	103	176	8	21
144	50.6	122	177	8.5	27
145	19.1	40	178	50.1	109
146	53	114	179	6.8	18
147	18	82	180	19.9	55
148	44.4	105	181	17.5	47
149	10.8	35	182	6.8	21
150	51.7	219	183	10.9	33
151	22.6	48	184	11.2	23
152	7.7	19	185	20.2	38
153	43.5	60	186	19.6	26
154	3.1	3	187	18.4	46
155	5	13	188	50.9	84
156	4.4	8	189	17.6	42
157	3.3	5	190	44.1	113
158	2.6	5	191	17	31
159	53.5	211	192	46.9	135
160	48.9	206	193	2.8	6
161	47.8	176	194	25.5	40
162	17.2	37	195	14.5	28
163	28.6	45	196	14.1	40
164	10.8	31	197	47.1	85
165	50.1	212	198	42.2	93

199	40.2	75	232	17.2	24
200	66.8	223	233	57	213
201	4.1	11	234	6.3	9
202	60.6	180	235	44.2	216
203	8	15	236	3	4
204	17.2	43	237	36.4	62
205	22	46	238	2.7	3
206	15.9	39	239	4.4	7
207	3.1	4	240	41.4	177
208	4.5	12	241	3.4	7
209	32	65	242	8.4	25
210	46.9	126	243	4.8	12
211	36.4	103	244	4.2	5
212	25.4	64	245	6.3	16
213	40	82	246	32.6	67
214	40.4	87	247	15.3	31
215	19.8	42	248	38.6	42
216	30.5	37	249	5.2	6
217	37.7	183	250	61.8	239
218	22.1	33	251	10.9	33
219	5.5	6	252	3.5	6
220	28.4	76	253	2.5	4
221	46.4	120	254	10.9	26
222	15.8	33	255	8.9	24
223	45.9	202	256	21	67
224	33.5	82	257	44.1	107
225	36.7	77	258	7	16
226	44	105	259	9.4	27
227	51.6	197	260	8	17
228	45	78	261	23	59
229	34	99	262	11.6	35
230	53.1	198	263	33	90
231	30.8	85	264	7.5	17

265	17.5	46	298	19.8	33
266	8.9	33	299	34	42
267	47.4	53	300	4.9	6
268	22	49	301	8.3	14
269	6.8	18	302	3.7	8
270	7.5	18	303	32.7	53
271	22.2	32	304	2.6	7
272	19.3	25	305	44.8	140
273	14.5	22	306	10.3	21
274	3.5	5	307	28.5	32
275	10.9	26	308	34	119
276	14.7	33	309	36.6	81
277	12.5	34	310	50.8	106
278	18.7	35	311	29.2	78
279	20.5	38	312	8.5	21
280	11.5	26	313	23.4	35
281	43.7	92	314	7.9	15
282	10.1	36	315	44.6	149
283	42.1	70	316	2.5	4
284	41.8	92	317	9.4	17
285	21.9	70	318	3	6
286	56.9	113	319	2.8	3
287	40.5	83	320	3	5
288	15.9	76	321	4.1	8
289	18.8	58	322	23.4	42
290	26.5	89	323	59	189
291	42.2	133	324	5.2	8
292	39.8	196	325	8.5	10
293	48.2	197	326	7.8	15
294	25.5	40	327	44.9	140
295	19.6	40	328	54.4	104
296	59.4	176	329	47.9	129
297	9.3	25	330	41.3	94

331	38.8	91	364	27	29
332	41.1	105	365	19.9	24
333	39	116	366	17.5	22
334	45.4	140	367	62.5	232
335	47.9	137	368	44	92
336	53.7	105	369	38	167
337	43.5	96	370	3.2	2
338	18.7	68	371	13.4	21
339	57.8	188	372	5.7	14
340	14.9	23	373	3.6	5
341	4.5	10	374	2.6	3
342	8.8	22	375	75.4	244
343	23.6	26	376	2.2	5
344	11.5	21	377	3.7	7
345	20	27	378	3.1	4
346	8.3	19	379	7.2	26
347	12.6	30	380	8.2	20
348	5.8	14	381	3.2	5
349	12.9	25	382	2.5	3
350	5.4	11	383	4	11
351	22.5	42	384	1.8	1
352	11.8	32	385	2.7	3
353	51.2	203	386	9.9	21
354	45.3	85	387	6.3	11
355	48.7	120	388	3.2	11
356	6.6	20	389	3.3	5
357	16.7	33	390	5	12
358	12.3	23	391	3.7	6
359	6.5	15	392	2	5
360	53	106	393	5.1	13
361	18.1	21	394	6	12
362	2.4	5	395	3.8	8
363	5.8	11	396	3.5	9

Tree Data Reprinted from Chen et al. (2004)

APPENDIX B:

R CODE

5.3 Functions

```
##Find Factors of a Number x
fac=function(x){
    div=seq(1,x)
    return(div[x%%div==0])
}

#### Uniform KPRSS Perfect Judgement Ranking####

ukprss_data=function(n,g,num,distr){ #g= number of groups, n=SRS size, num=total sample size

    data=matrix(nrow=num,ncol=g)

    for (i in 1:num){

        if(n==g){
            data[i,]=sort(distr(n))
        }
        else{
            this.data=sort(distr(n))
            for (G in 1:g){
                data[i,G]=this.data[sample((1+(G-1)*n/g):(G*n/g),size=1)]
            }
        }
    }
    return(data)
}


#### Balanced KPRSS Data ####
#k = size of k tuple
#g = group size
#n = SRS size
#M = number of cycles
#N = overall sample size
#sim = True: simulate from a known distribution
```

```
#sim = False: generate sample from a provided data frame
#distribution = random generation function for distribution. Ex rnorm, rbinom, rpois
#data_frame = data frame to generate samples from.
#... extra arguments for distribution function

bktup_data=function(k,g,n,M,sim=T,distribution=NA,data_frame=NA,...){
    labels=t(combn(n/g,k))
    cycles=array(dim=c(dim(labels)[1],k,M))
    for(m in 1:M){
        for(i in 1:dim(labels)[1]){
            if(sim==T){
                this.data=sort(do.call(distribution,list(n,...)))
            }else{
                this.data=sort(data_frame[sample(1:length(data_frame),size=n,replace=F)])
            }

            for(G in 1:dim(labels)[2]){
                if(g==1){
                    if(k==1){
                        cycles[i,G,m]=this.data[i]
                    }
                    if(k==n){
                        cycles[i,,m]=this.data
                    }

                }else{
                cycles[i,G,m]=this.data[sample((1+(labels[i,G]-1)*g):(labels[i,G]*g),size=1,
                replace=F)]
                }
            }
        }
    }

    return(list(data=cycles,labels=labels))
}

#### Balanced KPRSS using Concomitant Ranking Variable###
#k = size of k tuple
#g = group size
#n = SRS size
#M = number of cycles
#data_frame = data set
#voi = column from data set that is the variable of interest
#concom = column from data set to use for ranking


bktup_data2=function(k,g,n,M,data_frame=NA,voi=NA,concom=NA,...){
    labels=t(combn(n/g,k))
    cycles=array(dim=c(dim(labels)[1],k,M))
    ranking.error=rep(NA,dim(labels)[1]*M)
    for(m in 1:M){
        for(i in 1:dim(labels)[1]){
                sample.rows=data_frame[sample(1:dim(data_frame)[1],size=n,replace=F),]
                this.data=sample.rows[order(sample.rows[,concom]),]
```

```
                true.ranking=sort(sample.rows[,voi])
                ranking.error[(m-1)*choose(n/g,k)+i]=sum(true.ranking!=this.data[,voi])/n
            for(G in 1:dim(labels)[2]){
                if(g==1){
                    if(k==1){
                        cycles[i,G,m]=this.data[i,voi]
                    }
                    if(k==n){
                        cycles[i,,m]=this.data[,voi]
                    }
                    #print(cycles[i,G,m])
                }else{
                cycles[i,G,m]=this.data[sample((1+(labels[i,G]-1)*g):(labels[i,G]*g),
                size=1,replace=F),voi]
                }
            }
        }
    }

    return(list(data=cycles,labels=labels,ranking.error=ranking.error))
}
```

5.4 Simulation Code

5.4.1 For Section 2.2.1

Standard Normal

```
n=6 #SRS Size
k=c(1,2,3,6)
total=24 #Total Sample Size

results=matrix(nrow=length(k),ncol=7)
colnames(results)=c('NumberOfGroups','SRSSize','xbar','xkp','Var(Xbar)',
'Var(Xkp)','Var(xbar)/Var(xkp)')
for (j in k){
   reps=1000000
   xbar=rep(0,times=reps)
   xkp=rep(0,times=reps)
   num=total/j

   #####Normal#####
   for (i in 1:reps){
      this.data=ukprss_data(n=n,g=j,num=num,distr=function(x){rnorm(x)})
      xkp[i]=mean(this.data)
      xbar[i]=mean(rnorm(n=total)) #to check as an indicator that code is working
   }
   results[which(k==j),]=c(j,n,mean(xbar),mean(xkp),var(xbar),var(xkp),var(xbar)/var(xkp))
}
```

Gamma Scale = 1, Shape = 2

```
n=6 #SRS Size
k=c(1,2,3,6)
total=24 #Total Sample Size
results=matrix(nrow=length(k),ncol=7)
colnames(results)=c('k','SRSSize','xbar','xkp','Var(Xbar)','
Var(Xkp)','Var(xbar)/Var(xkp)')

for (j in k){
    reps=10000000
    xbar=rep(0,times=reps)
    xkp=rep(0,times=reps)
    num=total/j

    #####Normal#####
    for (i in 1:reps){
        this.data=ukprss_data(n=n,g=j,num=num,distr=function(x){rgamma(x,shape=2)})
        xkp[i]=mean(this.data)
        xbar[i]=mean(rgamma(n=total,shape=2))  #to check an indicator that code is working
    }
    results[which(k==j),]=c(j,n,mean(xbar),mean(xkp),var(xbar),var(xkp),var(xbar)/var(xkp))
}
```

5.4.2 For Section 3.2.1

Standard Normal

```
n=12
g=fac(n)
rep_num=10000

for(gg in g){
   ks=seq(1,n/gg,1)
   for(k in ks){
      this.result=rep(NA,times=rep_num)
      print(k)
      print(60/(k*choose(n/gg,k)))
      M=60/(k*choose(n/gg,k))
      if(M%%1==0){
         for(reps in 1:rep_num){
            this.data=bktup_data(k=k,n=n,g=gg,sim=T,
            M=60/(k*choose(n/gg,k)),distribution='rnorm')
            this.result[reps]=mean(this.data$data)
         }

      mean(this.result)
      var(this.result)
      }
   }
}
```

Tree Data

Simulated Perfect Judgement Ranking

```
n=12
g=fac(12)
N=60

for(gg in g){
   ks=seq(1,n/gg,1)
   for(k in ks){
      this.result=rep(NA,times=rep_num)
      M=N/(k*choose(n/gg,k))
      if(M%%1==0){
         for(reps in 1:rep_num){
         this.data=bktup_data(k=k,n=n,g=gg,M=N/(k*choose(n/gg,k)),
         N=N,data_frame=data[,2],sim=F)
         this.result[reps]=mean(this.data$data)
         }
         mean(this.result)
         var(this.result)
         }
   }
}
```

Ranking with Concomitant Variable

```
n=12
g=fac(12)
N=60

for(gg in g){
   ks=seq(1,n/gg,1)
   for(k in ks){
      this.result=rep(NA,times=rep_num)
      this.rankerror=c()
      M=N/(k*choose(n/gg,k))
      if(M%%1==0){
         for(reps in 1:rep_num){
         this.data=bktup_data2(k=k,n=n,g=gg,M=N/(k*choose(n/gg,k)),
         N=N,data_frame=data,voi=2,concom=1)
         this.result[reps]=mean(this.data$data)
         this.rankerror=c(this.rankerror,this.data$ranking.error)
         }
         mean(this.result)
         var(this.result)
         mean(this.rankerror)
      }
   }
}
```

5.4.3 For Section 4.3

```
n=12
g=c(1,fac(n))
N=50
rep_num=10000
xvals=qnorm(c(.05,.15,.25,.5,.75,.85,.95))

set.seed(81818)
for(gg in g){
   these.alphas=c(1:gg,(n-gg+1):n)
   these.betas=n-these.alphas+1
   hF=function(x,alpha,beta){
      val=sum(pbeta(x,alpha,beta))/(2*gg)
      return(val)
   }
   these.results=matrix(nrow=rep_num,ncol=length(xvals))
   for(rep in 1:rep_num){
      this.data=matrix(rep(NA,times=N),ncol=2)
      for(M in 1:N/2){
         this.draw=matrix(sort(rnorm(n)),ncol=gg,byrow=T)
         if(gg==1){
            this.data[M,]=c(this.draw[1,1],this.draw[n,1])
         }else{
         this.data[M,]=c(sample(this.draw[1,],1),sample(this.draw[dim(this.draw)[1],],1))
         }
      }
      this.ecdf=ecdf(this.data)
      for(y in 1:length(xvals)){
         these.results[rep,y]=bisect(0,1,func=function(x){hF(x,these.alphas,these.betas)
         -this.ecdf(xvals[y])},tol=.0000001)
      }
   }
   print(c(gg,colMeans(these.results)))
   print(c(gg,apply(X=these.results,MARGIN=2,FUN=var)))

}
```

BIBLIOGRAPHY

Aragon, M. E. D., Patil, G., and Taillie, C. (1999). A performance indicator for ranked set sampling using ranking error probability matrix. *Environmental and Ecological Statistics*, 6(1), 75–89.

Asghari, S., Gildeh, B. S., Ahmadi, J., and Borzadaran, G. M. (2017). Lifetime performance index based on ranked set sampling. *Communications in Statistics - Simulation and Computation*, 46(9), 7405–7422.

Bickel, P. J. (1967). Some Contributions To The Theory of Order Statistics. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 1, 575–591.

Chen, H., Stasny, E. A., and Wolfe, D. A. (2007). Improved procedures for estimation of disease prevalence using ranked set sampling. *Biometrical Journal*, 49(4), 530–538.

Chen, Z., Bai, Z., and Sinha, B. K. (2004). *Lecture Notes in Statistics 176: Ranked Set Sampling Theory and Applications*. Springer.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Dell, T. and Clutter, J. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28(2), 545–555.

Ghosh, K. and Tiwari, R. C. (2008). Estimating the distribution function using k-tuple ranked set samples. *Journal of Statistical Planning and Inference*, 138(4), 929–949.

Ghosh, K. and Tiwari, R. C. (2009). A unified approach to variations of ranked set sampling with applications. *Journal of Nonparametric Statistics*, 21(4), 471–485.

Haq, A., Brown, J., and Moltchanova, E. (2016). Hybrid ranked set sampling scheme. *Journal of Statistical Computation and Simulation*, 86(1), 1–28.

Harter, H. L. and Balakrishnan, N. (1996). *CRC Handbook of Tables for the use of Order Statistics in Estimation*. CRC Press, Boca Raton, Fl.

MacEachern, S. N., Öztürk, Ö., Wolfe, D. A., and Stark, G. V. (2002). A new ranked set sample estimator of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 177–188.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3(4), 385.

Mode, N. A., Conquest, L. L., and Marker, D. A. (1999). Ranked set sampling for ecological research: Accounting for the total costs of sampling. *Environmetrics*, 10(2), 179–194.

Muttlak, H. A. (1996). Pair Rank Set Sampling. *Biometrical Journal*, 38(7), 879–885.

Nahhas, R. W., Wolfe, D. A., and Chen, H. (2002). Ranked Set Sampling : Cost and Optimal Set Size. *Biometrics*, 58(4), 964–971.

Nazari, S. and Jozani, M. J. (2014). Nonparametric density estimation using partially rank-ordered set samples with application in estimating the distribution of wheat yield. *Electronic Journal of Statistics*, 8(1), 738–761.

Ozturk, O. (2010). Nonparametric maximum-likelihood estimation of within-set ranking errors in ranked set sampling. *Journal of Nonparametric Statistics*, 22(7), 823–840.

Ozturk, O. (2011). Sampling from partially rank-ordered sets. *Environmental and Ecological Statistics*, 18(4), 757–779.

Ozturk, O. and MacEachern, S. N. (2007). Order restricted randomized designs and two sample inference. *Environmental and Ecological Statistics*, 14(4), 365–381.

Platt, W. J., Evans, G. W., and Rathbun, S. L. (1988). The Population Dynamics of a Long-Lived Conifer (Pinus palustris). *The American Naturalist*, 131(4), 491–525.

Wang, Y.-G., Chen, Z., and Liu, J. (2004). General Ranked Set Sampling with Cost Considerations. *Biometrics*, 60(2), 556–561.

Wolfe, D. a. (2012). Ranked Set Sampling: Its Relevance and Impact on Statistical Inference. *ISRN Probability and Statistics*, 2012, 1–32.

Yu, P. L. H. and Lam, K. (1997). Regression Estimator in Ranked Set Sampling. *Biometrics*, 53(3), 1070–1080.

Zheng, G., Ghosh, K., Chen, Z., and Li, Z. (2006). Extreme rank selections for linkage analysis of quantitative trait loci using selected sib-pairs. *Annals of Human Genetics*, 70(6), 857–866.

CURRICULUM VITAE

Graduate College
University of Nevada, Las Vegas

Marvin Chris Javier
mcjavier@gmail.com

Degrees:

Bachelor of Arts in Mathematics and Molecular and Cell Biology, Dec. 2007
University of California, Berkeley

Master of Science in Statistics, May 2013
San Diego State University

Dissertation Title:

k-Tuple Sampling from Partially Rank-Ordered Sets

Dissertation Examination Committee:

Chairperson, Kaushik Ghosh, Ph.D.
Committee Member, Amei Amei, Ph.D.
Committee Member, Hokwon Cho, Ph.D.
Graduate Faculty Representative, Guogen Shan, Ph.D.

Work Experience

UCSD Autism Center of Excellence, Assistant Statistician, Oct 2011 -Aug 2012
Constructed generalized additive models for comparing brain development of autism spectrum children
to normal children.

Presentations

"The Practice of Rank Set Sampling", Nevada ASA Fall Symposium, UNLV, 2018

Publications:

Uniform k-Tuple Partial Rank-Ordered Set Sampling, In Preparation
Balanced k-Tuple Partial Rank-Ordered Set Sampling, In Preparation
General k-Tuple Partial Rank-Ordered Set Sampling, In Preparation