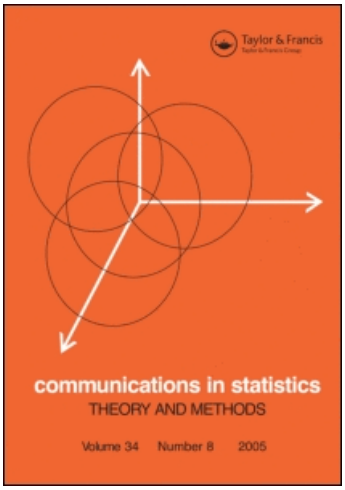

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

The Effectiveness of Multistage Ranked Set Sampling in Stratifying the Population

Monjed Hisham Samuh[a]; Mohammad Fraiwan Al-Saleh[b]
[a] Department of Mathematics and Computer Sciences, Palestine Polytechnic University, Hebron, Palestinian Territory, Occupied [b] Statistics Department, Yarmouk University, Jordan

Online publication date: 13 January 2011

To cite this Article Samuh, Monjed Hisham and Al-Saleh, Mohammad Fraiwan(2011) 'The Effectiveness of Multistage Ranked Set Sampling in Stratifying the Population', Communications in Statistics - Theory and Methods, 40: 6, 1063 — 1080
To link to this Article: DOI: 10.1080/03610920903521925
URL: http://dx.doi.org/10.1080/03610920903521925

Communications in Statistics—Theory and Methods, 40: 1063–1080, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920903521925

Taylor & Francis
Taylor & Francis Group

The Effectiveness of Multistage Ranked Set Sampling in Stratifying the Population

MONJED HISHAM SAMUH[1] AND
MOHAMMAD FRAIWAN AL-SALEH[2]

[1]Department of Mathematics and Computer Sciences,
Palestine Polytechnic University, Hebron, Palestinian Territory,
Occupied
[2]Statistics Department, Yarmouk University, Jordan

The effectiveness of Ranked Set Sampling (RSS) and Multistage Ranked Set Sampling (MSRSS) in stratifying the population is demonstrated. The overlapping coefficient of Weitzman (1970) is used as a numerical index to measure the degree of stratification at each stage. It turns out that this measure, under the assumption of no error in ranking, doesn't depend on the underlying distribution. Possible uses of this measure at the design stage and at the inference stage are addressed and briefly demonstrated. The overlapping coefficient in the case of judgment order statistics and its possible use in testing for perfect ranking is briefly addressed and put forward for future investigation.

Keywords Judgment order statistic; Multistage ranked set sampling; Overlapping coefficient; Ranked set sampling.

Mathematics Subject Classification 62G05; 62D05; 62H12.

1. Introduction

Ranked Set Sampling (RSS) was introduced by McIntyre (1952), (republished in McIntyre, 2005). To obtain a sample using this procedure: m sets of size m elements each are randomly selected from the population of interest. Then, for $i = 1, \ldots, m$, the ith minimum of the ith set is identified by judgment. The set of the m elements obtained is called a ranked set sample. When judgment ranking is accurate, these elements are actually a set of m *independent* order statistics. In practice, m should be small; a sample of size $n = hm$ can be obtained by performing the procedure h times.

Let the sample mean of RSS be denoted by $\hat{\mu}_{RSS}$ while the sample mean of a simple random sample (SRS) of the same size be denoted by $\hat{\mu}_{SRS}$. Mathematical theory of RSS was established by Takahasi and Wakimoto (1968). The authors showed that $\hat{\mu}_{RSS}$ is an unbiased estimator of the population mean, and has smaller

Received November 12, 2008; Accepted December 1, 2009
Address correspondence to Monjed Hisham Samuh, Department of Mathematics and Computer Sciences, Palestine Polytechnic University, Hebron, Palestinian Territory, Occupied; E-mail: mhstat@ppu.edu

variance than $\hat{\mu}_{SRS}$:

$$1 \leq eff(\hat{\mu}_{RSS}; \hat{\mu}_{SRS}) = \frac{Var(\hat{\mu}_{SRS})}{Var(\hat{\mu}_{RSS})} \leq \frac{m+1}{2}.$$

Al-Saleh and Al-Kadiri (2000) considered double RSS (DRSS), as a procedure that increases the efficiency of the RSS estimator without increasing the set size m. The procedure was generalized to Multistage RSS (MSRSS) by Al-Saleh and Al-Omari (2002). These procedures tend to produce an even more spread out data than RSS. The following steps describe MSRSS scheme.

1. m^{r+1} sample units are randomly selected from the target population, where r is the number of stages, and m is the set size; then allocated randomly into m^{r-1} sets, of size m^2 each.
2. For each set in Step 1, the RSS procedure described above is applied to obtain a (*judgment*) ranked set. This step yields m^{r-1} (*judgment*) ranked sets, of size m each.
3. Without doing any actual quantification on these ranked sets, repeat Step 2 on the m^{r-1} ranked sets to obtain m^{r-2} second stage (*judgment*) ranked sets, of size m each.
4. The process is continued using Step 3, without doing any actual quantification, until we end up with one rth stage ranked set of size m.
5. Repeat Steps 1–4 h times, if necessary, to obtain an rth stage RSS of size $n = mh$.

For illustration examples of MSRSS procedure, see Al-Saleh and Al-Omary (2002).

Since judgment ranking with large set size is prone to ranking errors, in practice, m should be small (2, 3, or 4). As given in the literature (see Takahasi and Wakimoto, 1968), for some distributions, such as highly skewed ones, small m doesn't lead to significant efficiency gain. One advantage of the MSRSS procedure is that efficiency gain can be achieved through the increase of the number of stages r rather than the set size m. It is a useful procedure to use as a tool of data reduction in the case of massive data sets; see Al-Saleh and Samuh (2008).

It should be mentioned here that in practice error in ranking is rarely avoidable. Thus, we rarely end up with a RSS that consists of truly independent order statistics. It should be also mentioned that RSS procedure tends to produce a more "spread out" set of data than that obtained using SRS procedure. Thus, ranking in the next stages tends to be easier than in the first stage (see Al-Saleh and Al-Kadiri, 2000). Unfortunately, the distributions of judgment order statistics are not easily derived. If the ranking on one variable is based on an auxiliary variable that is known to be correlated with the variable of interest, then the resulting judgment order statistics are actually concomitant order statistics. The distribution of concomitant order statistics can be derived (see Sec. 3).

For more work on RSS and its variations, see Chen et al. (2003), Al-Saleh (2004), Zheng and Al-Saleh (2002), Al-Saleh and Samawi (2007, 2009), Al-Saleh and Ananbeh (2007), Al-Saleh and Diab (2009), Oztork (2010), and Samawi et al. (2009).

RSS is closely related to stratified random sampling. In stratified sampling, the finite population of interest is divided into m nonoverlapping strata and a random sample of size h elements (in case of equal allocation) is taken from each stratum. The ideal situation occurs if the resulting strata are very different in between, but

similar within: the variability within each stratum is small compared to that between any two strata. On the other hand, in RSS the stratification occurs at the level of the sample; the sample itself is stratified. h elements are obtained randomly from the population of the ith (judgment) order statistic, for $i = 1, \ldots, m$. The $m$subpopulations are not truly non overlapping strata. The overlapping of any two subpopulations can be measured using the overlapping coefficient Δ. Throughout the article, $h = 1$.

For two probability density functions, f_1 & f_2, the overlapping coefficient, (Δ), of Weitzman (1970) is defined by

$$\Delta = \int_{-\infty}^{\infty} \min(f_1(x),\ f_2(x))dx.$$

Using the identity:

$$\min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$$

Δ, can be rewritten as

$$\Delta = 1 - \frac{1}{2}\int_{-\infty}^{\infty} |f_1(x) - f_2(x)|dx.$$

Figure 1 shows the overlapping coefficient (Δ) between two exponential distributions.

Al-Saleh (2007) used the overlapping coefficient (Δ) of Weitzman (1970), to measure the similarity between any pair of order statistics. It was shown that the similarity between any two consecutive order statistics $X_{(i)}$ and $X_{(i+1)}$ of a random sample $X_1, X_2, \ldots, X_n$ is

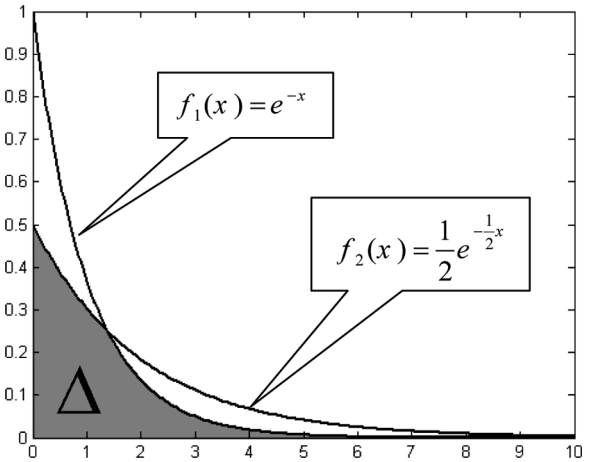

Figure 1. The overlapping coefficient for two exponential distributions.

$$\Delta(X_{(i)}, X_{(i+1)}) = 1 - \binom{n}{i}\left(\frac{i}{n}\right)^{i}\left(1 - \frac{i}{n}\right)^{n-i}.$$

and the similarity between the two extreme order statistics is

$$\Delta(X_{(1)}, X_{(n)}) = \left(\frac{1}{2}\right)^{n-1}.$$

Clearly, as $n \to \infty$, $\Delta(X_{(1)}, X_{(n)}) \to 0$; which means that the distribution of $X_{(1)}$ and $X_{(n)}$ tends not to overlap; in other words, the two distributions (or populations) tend to become two disjoint strata.

Many interpretations and possible uses of Δ have been mentioned in the literature. Clemons and Bradley (2000) proposed an interpretation of Δ based on classification of individuals into two populations: For an individual I, $\Delta = \Pr$ (I is classified in population 1 | I is actually from population 2) + Pr (I is classified in population 2 | I is actually from population 1). Thus, Δ can be regarded as the probability of misclassification. Actually, Δ is the size of what may be called "Indifference Zone". The above interpretation can be used to obtain an estimate of Δ based on a sample from the population of interest (see Sec. 3).

For more work on the overlapping coefficient, see Reiser and Faraggi (1999) and Clemons and Bradley (2000).

The main objective of this article is to use Δ to quantify the level of overlapping between the consecutive pairs of elements of multistage RSS (MSRSS). MSRSS eventually partitions the population into almost nonoverlapping portions (Strata); but through the value of Δ, the amount of overlapping is quantified at each stage. The computation of Δ in MSRSS is discussed in Sec. 2. Possible uses of this measure at the design and inference stage are addressed and briefly investigated, the overlapping coefficient in the case of error in ranking is also discussed in this section. Concluding remarks are given in Sec. 4.

2. The Overlapping Coefficient Based on MSRSS

In this section, the overlapping coefficient is obtained between any two densities of two consecutive data points of the MSRSS at different stages. It is assumed here that judgment ranking is as good as actual ranking; thus at each stage the resulting sample consists of independent order statistics. The more realistic case, in which error in ranking exists, will be dealt with in Sec. 3.

Assume that the variable of interest has a probability density function (pdf) f, with absolutely continuous cumulative distribution function (cdf) F; Let $Y_i^{(r)}$, $(i = 1, 2, \ldots, m)$, be the ith element of an MSRSS of size m at stage r. Thus, for $r = 1, 2, ...,$ $Y_i^{(r)}$ is the ith judgment minimum of $Y_1^{(r-1)}, Y_2^{(r-1)}, \ldots, Y_m^{(r-1)}$. Let $f_i^{(r)}$ be its pdf and $F_i^{(r)}$ be its cdf. Under the assumption of no error in judgment ranking, $f_i^{(r)}$ is the density of the ith order statistic of a MSRSS, $Y_1^{(r-1)}, Y_2^{(r-1)}, \ldots, Y_m^{(r-1)}$, i.e., $Y_i^{(r)} =^d Y_{(i)}^{(r-1)}$; $Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_m^{(r)}$ are independent. The independence follows from the fact that in RSS (and MSRSS) each measured element is taken based on independent samples (see step 1 above).

Let $Y_i^{(r)}, Y_{i+1}^{(r)}$ be any two consecutive data points of a MSRSS at stage r, $i = 1, 2, \ldots, m-1$. Then, the overlapping coefficient between their probability density functions is

$$\Delta_{(i,i+1)}^{(r)} = \int_{-\infty}^{\infty} \min\left(f_{i+1}^{(r)}(x),\ f_i^{(r)}(x)\right) dx,$$

where

$$F_k^{(r)}(x) = \sum_{t=k}^{m} \sum_{S_t} \prod_{k=1}^{t} F_{j_k}^{(r-1)}\left(\frac{i}{m}\right) \prod_{k=t+1}^{m} \left(1 - F_{j_k}^{(r-1)}\left(\frac{i}{m}\right)\right);$$

S_t is the set of all permutations, $(i_1, i_2, i_3, \ldots, i_m)$, of the numbers $(1,2,\ldots,m)$ for which $i_1 < \cdots < i_t$ and $i_{t+1} < i_{t+2} < \cdots < i_m$. (Al-Saleh and Al-Omari, 2002).

Note that $F_k^{(r)}(x)$ is a function of $F(x)$, $F_k^{(r)}(x) = H_k^{(r)}(F(x))$, say.

Hence,

$$\Delta_{(i,i+1)}^{(r)} = \int_{-\infty}^{\infty} \min\left(f(x)h_{i+1}^{(r)}(F(x)),\ f(x)h_i^{(r)}(F(x))\right) dx,$$

where $h_k^{(r)}(x) = d/dx H_k^{(r)}(x)$. Now, $u = F(x) \Rightarrow du = f(x)dx$; thus,

$$\Delta_{(i,i+1)}^{(r)} = \int_0^1 \min\left(h_{i+1}^{(r)}(u), h_i^{(r)}(u)\right) du.$$

Thus, $\Delta_{(i,i+1)}^{(r)}$ doesn't depend on $F(x)$. Hence, without loss of generality, we may assume that we are sampling from the uniform distribution; $f(x) = 1$, for $0 < x < 1$ and zero otherwise.

2.1. *Special Cases*

1. Consider the set size $m = 2$; that is, we have

$$Y_1^{(r)} \sim F_1^{(r)}(x) = 1 - \left(1 - F_1^{(r-1)}(x)\right)\left(1 - F_2^{(r-1)}(x)\right)$$

and

$$Y_2^{(r)} \sim F_2^{(r)}(x) = 2F(x) - F_1^{(r)}(x).$$

$$\Delta_{(1,2)}^{(r)} = \int_{-\infty}^{\infty} \min\left(f_1^{(r)}(x),\ f_2^{(r)}(x)\right) dx.$$

It can be verified that

$$\min\left(f_1^{(r)}(x),\, f_2^{(r)}(x)\right) = \begin{cases} 2f(x) - f_2^{(r)}(x), & f(x) < f_2^{(r)}(x) \\ f_2^{(r)}(x), & f(x) \geq f_2^{(r)}(x) \end{cases}.$$

- For $r = 1$;

$$f_2^{(1)}(x) = 2f(x)F(x),$$

hence,

$$\min\left(f_1^{(1)}(x),\, f_2^{(1)}(x)\right) = \begin{cases} 2f(x)(1 - F(x)), & F(x) > 0.5 \\ 2f(x)F(x), & F(x) \leq 0.5 \end{cases}.$$

Thus,

$$\Delta_{(1,2)}^{(1)} = \int_{0.5}^{1} 2(1-u)du + \int_{0}^{0.5} 2u\,du = 0.5.$$

- For $r = 2$;

$$f_2^{(2)}(x) = 6f(x)F^2(x) - 4f(x)F^3(x)$$

hence,

$$\min\left(f_1^{(2)}(x),\, f_2^{(2)}(x)\right) = \begin{cases} 2f(x)\left(1 - 3F^2(x) + 2F^3(x)\right), & F(x) > 0.5 \\ 2f(x)\left(3F^2(x) - 2F^3(x)\right), & F(x) \leq 0.5 \end{cases}.$$

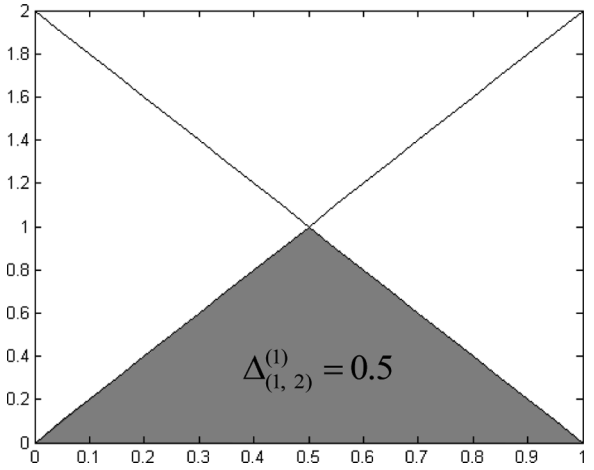

Figure 2. The overlapping coefficient when the set size $m = 2$ at stage $r = 1$.

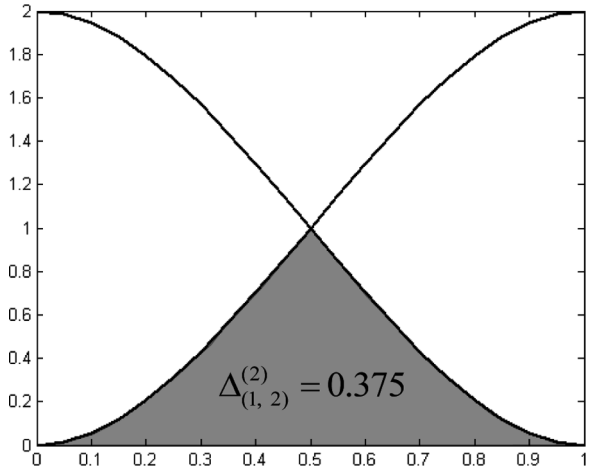

Figure 3. The overlapping coefficient when the set size $m = 2$ at stage $r = 2$.

Thus,

$$\Delta^{(2)}_{(1,2)} = \int_{0.5}^{1} 2\left(1 - 3u^2 + 2u^3\right)du + \int_{0}^{0.5} \left(6u^2 - 4u^3\right)du = 0.375.$$

Similarly, it can be shown that the overlapping coefficient between $Y_1^{(r)}$ and $Y_2^{(r)}$ at stage $r = 3, 4, 5$ are $\Delta^{(3)}_{(1,2)} = 0.30469$, $\Delta^{(4)}_{(1,2)} = 0.25827$, and $\Delta^{(5)}_{(1,2)} = 0.22492$, respectively. Clearly, $\Delta^{(r)}$ is decreasing in r. As $r \to \infty$, using Al-Saleh and Al-Omari (2002), $Y_1^{(\infty)} \sim F_1^{(\infty)}(x) = 2F(x)$, for $F(x) < 0.5$ and 1 otherwise; $f_1^{(\infty)}(u) = 2$, for $u < 0.5$ and zero otherwise; $Y_2^{(\infty)} \sim F_2^{(\infty)}(x) = 2F(x) - 1$, for $F(x) \geq 0.5$ and 1 otherwise; $f_2^{(\infty)}(u) = 2$, for $u \geq 0.5$ and zero otherwise. Thus, the overlapping coefficient between $Y_1^{(\infty)}$ & $Y_2^{(\infty)}$ is

$$\Delta^{(\infty)}_{(1,2)} = \int_{0}^{1} \min\left(f_1^{(\infty)}(u), f_2^{(\infty)}(u)\right)du = 0.$$

2. Consider the set size $m = 3$; that is, we have

$$Y_1^{(r)} \sim F_1^{(r)}(x) = 1 - \left(1 - F_1^{(r-1)}(x)\right)\left(1 - F_2^{(r-1)}(x)\right)\left(1 - F_2^{(r-1)}(x)\right),$$
$$Y_2^{(r)} \sim F_2^{(r)}(x) = 3F(x) - F_1^{(r)}(x) - F_3^{(r)}(x),$$

and

$$Y_3^{(r)} \sim F_3^{(r)}(x) = F_1^{(r-1)}(x)F_2^{(r-1)}(x)F_3^{(r-1)}(x).$$

For $r = 1$;

- The overlapping coefficient between $Y_1^{(1)}$ and $Y_2^{(1)}$

$$\min\left(f_1^{(1)}(x), f_2^{(1)}(x)\right) = \begin{cases} 3f(x)(1-F(x))^2, \\ \quad 3f(x)(1-F(x))^2 < 3f(x) - 3f(x)(1-F(x))^2 - 3f(x)F^2(x) \\ 3f(x)\left(1-(1-F(x))^2 - F^2(x)\right), & o.w \end{cases}$$

but,

$$3f(x)(1-F(x))^2 < 3f(x) - 3f(x)(1-F(x))^2 - 3f(x)F^2(x) \Leftrightarrow F(x) > \frac{1}{3}.$$

Thus,

$$\Delta_{(1,2)}^{(1)} = \int_{-\infty}^{\infty} \min\left(f_1^{(1)}(x),\, f_2^{(1)}(x)\right) dx = \int_{\frac{1}{3}}^{1} 3(1-u)^2 du + \int_0^{\frac{1}{3}} 3\left(1-(1-u)^2 - u^2\right) du = 0.55556.$$

- The overlapping coefficient between $Y_2^{(1)}$ and $Y_3^{(1)}$

$$\min\left(f_2^{(1)}(x), f_3^{(1)}(x)\right) = \begin{cases} 6f(x)\left(F(x) - F^2(x)\right), \\ \quad 6f(x)\left(F(x) - F^2(x)\right) < 3f(x)F^2(x) \\ 3f(x)F^2(x), & o.w \end{cases}$$

but,

$$6f(x)(F(x) - F^2(x)) < 3f(x)F^2(x) \Leftrightarrow F(x) > \frac{2}{3}.$$

Thus,

$$\Delta_{(2,3)}^{(1)} = \int_{-\infty}^{\infty} \min\left(f_2^{(1)}(x), f_3^{(1)}(x)\right) dx = \int_{\frac{2}{3}}^{1} 6\left(u - u^2\right) du + \int_0^{\frac{2}{3}} 3u^2 du = 0.55556.$$

Similarly, it can be verified that $\Delta_{(1,2)}^{(2)} = 0.4156$, $\Delta_{(2,3)}^{(2)} = 0.4156$, $\Delta_{(1,2)}^{(3)} = 0.3326$ and $\Delta_{(2,3)}^{(3)} = 0.3326$.

Based on empirical work, we have the following result that has not been proved for general m. We have only proved it for $m = 2$. It should b emphasized that Δ can be obtained directly as above without the conjecture given below.

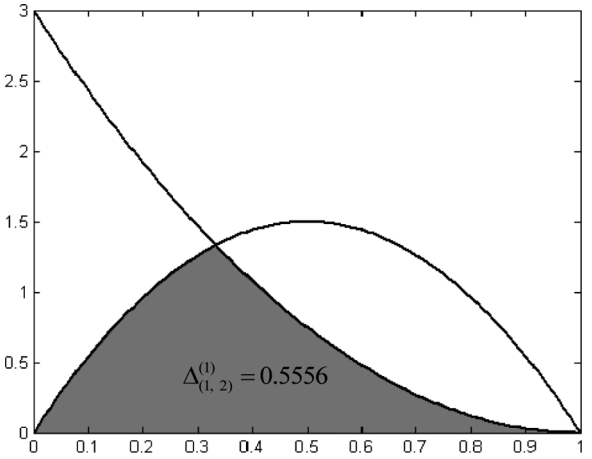

Figure 4. The overlapping coefficient between $Y_1^{(1)}$ and $Y_2^{(1)}$ when the set size $m = 3$ at stage $r = 1$.

Conjecture. *We claim that* $f_{i+1}^{(r)}(x) - f_i^{(r)}(x) = 0$ at $x = i/m$, *and*

$$\min\left(f_{i+1}^{(r)}(x),\, f_i^{(r)}(x)\right) = \begin{cases} f_{i+1}^{(r)}(x), & x < \frac{i}{m} \\ f_i^{(r)}(x), & x > \frac{i}{m} \end{cases}$$

This statement is proved for $m = 2$.

Proof. Let $g^{(r)}(x) = f_2^{(r)}(x) - f_1^{(r)}(x)$. We claim that $g^{(r)}(1/2) = 0$.
This statement is proved by mathematical induction.

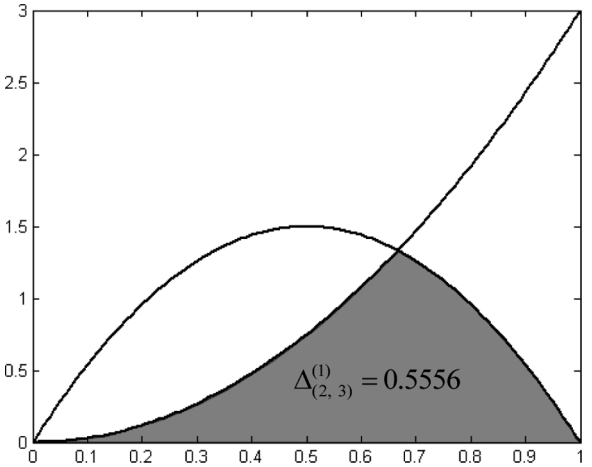

Figure 5. The overlapping coefficient between $Y_2^{(1)}$ and $Y_3^{(1)}$ when the set size $m = 3$ at stage $r = 1$.

For $r=1$,

$$f_2^{(1)}(x)=2x \quad \text{and} \quad f_1^{(1)}(x)=2(1-x).$$

Thus,

$$g^{(1)}\left(\frac{1}{2}\right)=f_2^{(1)}\left(\frac{1}{2}\right)-f_1^{(1)}\left(\frac{1}{2}\right)=0$$

Assume $g^{(r-1)}(1/2)=0$; that is, $f_2^{(r-1)}(1/2)-f_1^{(r-1)}(1/2)=0$.

We want to show that $g^{(r)}(1/2)=0$.

$$g^{(r)}\left(\frac{1}{2}\right)=f_2^{(r)}\left(\frac{1}{2}\right)-f_1^{(r)}\left(\frac{1}{2}\right)=f_2^{(r)}\left(\frac{1}{2}\right)-\left(2-f_2^{(r)}\left(\frac{1}{2}\right)\right)=2f_2^{(r)}\left(\frac{1}{2}\right)-2.$$

But,

$$\begin{aligned} f_2^{(r)}\left(\frac{1}{2}\right) &= (F_1^{(r-1)}(x)F_2^{(r-1)}(x))'\Big|_{x=\frac{1}{2}} \\ &= F_1^{(r-1)}\left(\frac{1}{2}\right)f_2^{(r-1)}\left(\frac{1}{2}\right)+F_2^{(r-1)}\left(\frac{1}{2}\right)f_1^{(r-1)}\left(\frac{1}{2}\right). \end{aligned}$$

Since $f_2^{(r-1)}(1/2)-f_1^{(r-1)}(1/2)=0$ is true, and using the fact $f(x)=1/m\sum_{i=1}^{m}f_i^{(r)}(x)$ (see Al-Saleh and Al-Omari, 2002), we have

$$f_2^{(r)}\left(\frac{1}{2}\right)=f_2^{(r-1)}\left(\frac{1}{2}\right)\left(F_1^{(r-1)}\left(\frac{1}{2}\right)+F_2^{(r-1)}\left(\frac{1}{2}\right)\right)=f_2^{(r-1)}\left(\frac{1}{2}\right).$$

Thus,

$$f_2^{(r-1)}\left(\frac{1}{2}\right)=f_2^{(r-2)}\left(\frac{1}{2}\right)=\cdots=f_2\left(\frac{1}{2}\right)=1.$$

Therefore,

$$g\left(\frac{1}{2}\right)=2f_2^{(r)}\left(\frac{1}{2}\right)-2=0.$$

Now, we want to show that

$$\min(f_2^{(r)}(x),f_1^{(r)}(x))=\begin{cases} f_2^{(r)}(x), & x<\frac{1}{2} \\ f_1^{(r)}(x), & x>\frac{1}{2} \end{cases}$$

which is equivalent to showing that $g^{(r)}(x)$ is an increasing function; that is, $g'^{(r)}(x)>0$.

For $r=1$,

$$g^{(1)}(x)=2f_2^{(1)}(x)-2f(x)=2(F(x)f(x)+f(x)F(x))-2f(x)=2f_2^{(1)}(x)-2=4x-2.$$

Thus, $g'^{(1)}(x)=4>0$; Therefore, $g^{(1)}(x)$ is an increasing function.

Assume $g^{(r-1)}(x)$ is increasing; that is, $g'^{(r-1)}(x)=2f_2'^{(r-1)}(x)>0$.

At stage r:

$$\begin{aligned}
g^{(r)}(x) &= 2(F_1^{(r-1)}(x)f_2^{(r-1)}(x)+f_1^{(r-1)}(x)F_2^{(r-1)}(x))-2,\\
g'^{(r)}(x) &= 2(F_1^{(r-1)}(x)f_2'^{(r-1)}(x)+f_2^{(r-1)}(x)f_1^{(r-1)}(x)\\
&\quad +f_1^{(r-1)}(x)f_2^{(r-1)}(x)+f_1'^{(r-1)}(x)F_2^{(r-1)}(x))\\
&= 2(2f_1^{(r-1)}(x)f_2^{(r-1)}(x)+f_2'^{(r-1)}(x)(F_1^{(r-1)}(x)-F_2^{(r-1)}(x)))\\
&= 2(2f_1^{(r-1)}(x)f_2^{(r-1)}(x)+f_2'^{(r-1)}(x)(2F(x)-2F_2^{(r-1)}(x))).
\end{aligned}$$

The last term in the above equation is positive provided that $F(x)-F_2^{(r-1)}(x)>0$. This is also shown by induction.

For $r=1$;

$$F(x)>F^2(x),\quad \text{for}\quad 0<x<1.$$

Assume that $F(x)-F_2^{(r-2)}(x)>0\Rightarrow F(x)>F_2^{(r-2)}(x)$.

At stage $r-1$:

$$F_2^{(r-1)}(x)=F_1^{(r-2)}(x)F_2^{(r-2)}(x)<F_2^{(r-2)}(x),$$

thus,

$$F>F_2^{(r-1)}(x).$$

Therefore, $g^{(r)}(x)$ is increasing function.

Since $g^{(r)}(x)$ is increasing function and it has a root at $x=1/2$, then

$$\min(f_2^{(r)}(x),\, f_1^{(r)}(x))=\begin{cases} f_2^{(r)}(x), & x<\frac{1}{2}\\ f_1^{(r)}(x), & x>\frac{1}{2}\end{cases}$$

Note. All numerical calculations of Δ agree with the calculations based on the above conjecture. Thus, we strongly believe that the result is true for general m; but have not been able to prove it. □

2.2. *General Recursive Formula for* $\Delta_{(i,i+1)}^{(r)}$

The overlapping coefficient between any two consecutive data points of a MSRSS at stage r, $Y_i^{(r)}$, $Y_{i+1}^{(r)}$, $i=1,2,\ldots,m-1$ is

$$\Delta_{(i,i+1)}^{(r)} = \int_{-\infty}^{\infty} \min\left(f_{i+1}^{(r)}(x),\, f_i^{(r)}(x)\right) dx.$$

When $m = 2$, we have:

$$\begin{aligned}
\Delta_{(1,2)}^{(r)} &= \int_0^{\frac{1}{2}} f_2^{(r)}(x)dx + \int_{\frac{1}{2}}^1 f_1^{(r)}(x)dx = \int_0^{\frac{1}{2}} f_2^{(r)}(x)dx + \int_{\frac{1}{2}}^1 2 - f_2^{(r)}(x)dx \\
&= 2F_2^{(r)}\left(\frac{1}{2}\right) = 2\left(F_1^{(r-1)}\left(\frac{1}{2}\right)F_2^{(r-1)}\left(\frac{1}{2}\right)\right) \\
&= 2\left(2F\left(\frac{1}{2}\right) - F_1^{(r-1)}\left(\frac{1}{2}\right)\right)F_2^{(r-1)}\left(\frac{1}{2}\right) \\
&= 2\left(1 - F_2^{(r-1)}\left(\frac{1}{2}\right)\right)F_2^{(r-1)}\left(\frac{1}{2}\right) \\
&= \left(1 - \frac{1}{2}\Delta_{(1,2)}^{(r-1)}\right)\Delta_{(1,2)}^{(r-1)}.
\end{aligned}$$

This result can be generalized for general m, using the conjecture:

$$\begin{aligned}
\Delta_{(i,i+1)}^{(r)} &= \int_0^{\frac{i}{m}} f_{i+1}^{(r)}(x)dx + \int_{\frac{i}{m}}^1 f_i^{(r)}(x)dx = F_{i+1}^{(r)}(x)\Big]_0^{\frac{i}{m}} + F_i^{(r)}(x)\Big]_{\frac{i}{m}}^1 \\
&= 1 - F_i^{(r)}\left(\frac{i}{m}\right) + F_{i+1}^{(r)}\left(\frac{i}{m}\right).
\end{aligned}$$

But,

$$F_i^{(r)}\left(\frac{i}{m}\right) = P\left(\text{at least } i \text{ of } \left(Y_1^{(r-1)},\ Y_2^{(r-1)}, \ldots, Y_m^{(r-1)}\right) < \frac{i}{m}\right),$$

and

$$F_{i+1}^{(r)}\left(\frac{i}{m}\right) = P\left(\text{at least } i+1 \text{ of } \left(Y_1^{(r-1)},\ Y_2^{(r-1)}, \ldots, Y_m^{(r-1)}\right) < \frac{i}{m}\right).$$

Therefore,

$$\begin{aligned}
\Delta_{(i,i+1)}^{(r)} &= 1 - P\left(\text{exactly } i \text{ of } \left(Y_1^{(r-1)},\ Y_2^{(r-1)}, \ldots, Y_m^{(r-1)}\right) < \frac{i}{m}\right) \\
&= 1 - \sum_{S_t} \prod_{k=1}^{t} F_{j_k}^{(r-1)}(i/m) \prod_{k=t+1}^{m} \left(1 - F_{j_k}^{(r-1)}(i/m)\right)
\end{aligned}$$

where S_t is the set of all permutations $(j_1, j_2, j_3, \ldots, j_m)$, of the numbers $(1, 2, \ldots, m)$ for which $j_1 < j_2 < j_3 < \cdots < j_i$, $j_{i+1} < j_{i+2} < j_{i+3} < \cdots < j_m$.

2.3. *Examples*

1. Consider the set size, $m = 2$.

   The overlapping coefficient between the density functions of $Y_1^{(1)}$ and $Y_2^{(1)}$, was shown to be $\Delta_{(1,2)}^{(1)} = 0.5$; then,

$$\Delta_{(1,2)}^{(2)} = \left(1 - \frac{1}{2}\Delta_{(1,2)}^{(1)}\right)\Delta_{(1,2)}^{(1)} = 0.37500,$$
$$\Delta_{(1,2)}^{(3)} = \left(1 - \frac{1}{2}\Delta_{(1,2)}^{(2)}\right)\Delta_{(1,2)}^{(2)} = 0.30469,$$
$$\Delta_{(1,2)}^{(4)} = \left(1 - \frac{1}{2}\Delta_{(1,2)}^{(3)}\right)\Delta_{(1,2)}^{(3)} = 0.25827,$$

   and so on.

2. Consider the set size, $m = 3$.

   The overlapping coefficient between the density functions of $Y_1^{(r)}$ and $Y_2^{(r)}$ is

$$\begin{aligned}\Delta_{(1,2)}^{(r)} &= 1 - P\left(\text{exactly 1 of}\left(Y_1^{(r-1)}, Y_2^{(r-1)}, \ldots, Y_m^{(r-1)}\right) < \frac{1}{3}\right)\\ &= 1 - \left(F_1^{(r-1)}\left(\frac{1}{3}\right)\left(1 - F_2^{(r-1)}\left(\frac{1}{3}\right)\right)\left(1 - F_3^{(r-1)}\left(\frac{1}{3}\right)\right)\right.\\ &\quad + F_2^{(r-1)}\left(\frac{1}{3}\right)\left(1 - F_1^{(r-1)}\left(\frac{1}{3}\right)\right)\left(1 - F_3^{(r-1)}\left(\frac{1}{3}\right)\right)\\ &\quad \left. + F_3^{(r-1)}\left(\frac{1}{3}\right)\left(1 - F_1^{(r-1)}\left(\frac{1}{3}\right)\right)\left(1 - F_2^{(r-1)}\left(\frac{1}{3}\right)\right)\right).\end{aligned}$$

- **For** $r = 1$:

$$\Delta_{(1,2)}^{(1)} = 1 - \left(3F\left(\frac{1}{3}\right)\left(1 - F\left(\frac{1}{3}\right)\right)^2\right) = 1 - \left(\frac{2}{3}\right)^2 = \frac{5}{9} = 0.55556.$$

- **For** $r = 2$:

$$F_1^{(1)}\left(\frac{1}{3}\right) = 1 - \left(1 - F\left(\frac{1}{3}\right)\right)^3 = \frac{19}{27} = 0.70370,$$
$$F_3^{(1)}\left(\frac{1}{3}\right) = F^3\left(\frac{1}{3}\right) = \left(\frac{1}{3}\right)^3 = \frac{1}{27} = 0.03704,$$

  and

$$F_2^{(1)} = 3F\left(\frac{1}{3}\right) - F_1^{(1)}\left(\frac{1}{3}\right) - F_3^{(1)}\left(\frac{1}{3}\right) = 1 - \frac{19}{27} - \frac{1}{27} = \frac{7}{27}.$$
$$\begin{aligned}\Delta_{(1,2)}^{(2)} &= 1 - \left(\left(\frac{19}{27}\right)\left(1 - \frac{7}{27}\right)\left(1 - \frac{1}{27}\right) + \left(1 - \frac{19}{27}\right)\left(1 - \frac{7}{27}\right)\left(\frac{1}{27}\right)\right.\\ &\quad \left. + \left(1 - \frac{19}{27}\right)\left(\frac{7}{27}\right)\left(1 - \frac{1}{27}\right)\right) = \frac{2729}{6561} = 0.41594,\end{aligned}$$

  and so on.

Table 1

$\Delta^{(r)}_{(i,j)}\, i = 1, 2, \ldots, m-1,\ m = 2, 3, 4,\ r = 1, 2, \ldots, 10$

$m\rightarrow$	2	3		4		
$r_{\downarrow}$	$\Delta^{(r)}_{(1,2)}$	$\Delta^{(r)}_{(1,2)}$	$\Delta^{(r)}_{(2,3)}$	$\Delta^{(r)}_{(1,2)}$	$\Delta^{(r)}_{(2,3)}$	$\Delta^{(r)}_{(3,4)}$
1	0.5000	0.5556	0.5556	0.5781	0.6250	0.5781
2	0.3750	0.4156	0.4156	0.4313	0.4713	0.4313
3	0.3047	0.3326	0.3326	0.3430	0.3717	0.3430
4	0.2583	0.2780	0.2780	0.2854	0.3051	0.2854
5	0.2249	0.2395	0.2395	0.2448	0.2589	0.2448
6	0.1996	0.2108	0.2108	0.2149	0.2255	0.2149
7	0.1797	0.1886	0.1886	0.1918	0.2001	0.1918
8	0.1636	0.1708	0.1708	0.1734	0.1800	0.1734
9	0.1502	0.1562	0.1562	0.1584	0.1638	0.1584
10	0.1389	0.1440	0.1440	0.1458	0.1504	0.1458
∞	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1 gives the overlapping coefficient between $Y_i^{(r)}$ and $Y_{i+1}^{(r)}$, at different stages when the set size $m = 2, 3, 4$.

Clearly, $\Delta^{(r)}_{(i,j)} \downarrow 0$ as $r \rightarrow \infty$. Thus, as the number of stages goes up, the more the sample elements tend to be from disjoint strata. It can be seen that $\Delta^{(r)}_{(i,i+1)} = \Delta^{(r)}_{(m-i,m-i+1)}$; thus, $\Delta^{(r)}$ need to be calculated for only half of the adjacent pairs. Also for any fixed r and (i,j), $\Delta^{(r)}_{(i,j)}$ is increasing in m. Thus, a certain degree of overlapping can be attained by choosing suitable r and m.

3. Possible Uses of the Overlapping Coefficient Δ

In the previous section, the overlapping coefficient Δ was calculated for each adjacent pairs of the elements of RSS and MSRSS, assuming that the set of elements are the actual set of order statistics, i.e., judgment order statistics are actual order statistics, a claim which is rarely true in practice. It was seen above that Δ doesn't depend on the underlying distribution; thus the value is known for each m and r. In practice, knowing the values of Δ can have several advantages and uses. The following are some of these possible uses of Δ at the statistical planning stage and statistical inference stage; some of them were hinted by the referees and an Associate Editor.

1. At the Planning Stage: Knowing the value of Δ in advance, independence of the underlying distribution, can be used for designing purposes. Being a function of m and r, Δ can help to make a decision on the value of the set size m and the value of the number of stages r. The very basic notion of RSS came from the idea of stratified sampling; thus the more the level of stratification (small Δ) the higher the efficiency of RSS in estimating the population parameters compared to SRS. But this efficiency depends on the underlying distribution while Δ doesn't. For

fixed adjacent pair, Δ is decreasing in r for fixed m, but increasing in m for fixed r; thus, the investigator needs to make a choice between increasing m, a decision that may increases efficiency but introduces judgment error, or increasing r, which means more efforts. Note that the ranking error tends to get smaller as r increases, this is intuitively clear because the elements at a stage are more spaced than the previous stage and hence easier to rank, (see Al-Saleh and Al-Kadiri, 2000). For example, if we want our sample to be so that the overlapping coefficient between consecutive strata is at most 30%, then we may take $(m, r) = (2, 3)$ or (4, 4), etc. This level of stratification can be achieved also with $r = 1$ but of course with much larger m.

2. At the Inference Stage: Once the data are collected from the chosen RSS or MSRSS, it is used to make inference about the population parameters. The accuracy of the estimators and their efficiencies with respect to the corresponding SRS estimators depend on their underlying distributions. One thing that has to be evaluated is the accuracy of the judgment ranking: Are the collected observations truly a set of independent order statistics? Is the error in judgment ranking too big so that the resulting chosen sample is just as good as SRS?

One way to address this problem is to see whether the overlapping coefficients between adjacent pairs in the chosen RSS or MSRSS sample are close to those given above for "error-free" judgment ranking. For example, with $m = 3$ & $r = 1$, $\Delta^{(1)}_{(1,2)} = \Delta^{(1)}_{(2,3)} = 0.5556$.

One way of obtaining a RSS or a MSRSS is to rank the characteristic of interest Y based or another auxiliary variable X which is easily ranked and correlated with Y. Let the joint density of (X, Y) be $f_{X,Y}(x, y)$ and f_X, f_Y be the marginal density of X and Y, respectively, and let $f_{Y|X}$ be the conditional density of Y given X. Let $Y_{[i]}$ denote the concomitant of $X_{(i)}$, then $Y_{[i]}$ can be regarded as a ith judgment order statistic. It was shown in Yang (1977) that

$$f_{Y_{[i]}|X_{(i)}}(y|x) = f_{Y|X}(y|x).$$

This result was generalized to MSRSS by Al-Saleh and Zheng (2002). Using this result, the density of $Y_{[i]}$ is

$$f_{Y_{[i]}}(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_{X_{(i)}}(x) dx.$$

Thus, the overlapping coefficient between the adjacent concomitant variables $Y_{[i]}$ and $Y_{[i+1]}$ is

$$\begin{aligned}\Delta^{*(1)}_{(i,i+1)} &= 1 - \frac{1}{2}\int_{-\infty}^{\infty} |f_{Y_{[i+1]}}(y) - f_{Y_{[i]}}(y))| dy \\ &= 1 - \frac{1}{2}\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} f_{Y|X}(y|x) \left(f_{X_{(i+1)}}(x) - f_{X_{(i)}}(x) \right) dx \right| dy.\end{aligned}$$

Also, from the above relation, using the triangular inequality, we have

$$\begin{aligned}\Delta^{*(1)}_{(i,i+1)} &\geq 1-\frac{1}{2}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f_{Y|X}(y|x)\left|f_{X_{(i+1)}}(x)-f_{X_{(i)}}(x)\right|dx\right)dy\\ &= 1-\frac{1}{2}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f_{Y|X}(y|x)\left|f_{X_{(i+1)}}(x)-f_{X_{(i)}}(x)\right|dy\right)dx\\ &= 1-\frac{1}{2}\int_{-\infty}^{\infty}\left|f_{X_{(i+1)}}(x)-f_{X_{(i)}}(x)\right|dx=\Delta^{(1)}_{(i,i+1)}\end{aligned}$$

Thus, $\Delta^{*(1)}_{(i,i+1)}$ depends on the underlying distribution and is always larger than or equal $\Delta^{(1)}_{(i,i+1)}$. It can be calculated for a given $f_{X,Y}(x, y)$, after estimating some of unknown parameters. The above theory can be easily generalized for MSRSS. For example, when $m=2$ and if $f_{X,Y}(x, y)$ is the density of a bivariate normal random variable, $BN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, it can be verified that

$$Y_{[1]} \sim f_{Y_{[1]}}(y) = 2\phi\left(\frac{y-\mu_y}{\sigma_y}\right)\left(1-\Phi\left(\frac{\rho}{\sqrt{2-\rho^2}}\frac{y-\mu_y}{\sigma_y}\right)\right),$$

and

$$Y_{[2]} \sim f_{Y_{[2]}}(y) = 2\phi\left(\frac{y-\mu_y}{\sigma_y}\right)\Phi\left(\frac{\rho}{\sqrt{2-\rho^2}}\frac{y-\mu_y}{\sigma_y}\right),$$

where ϕ and Φ are, respectively, the density and the cumulative distribution function of a standard normal random variable. Thus,

$$\Delta^{*(1)}_{(1,2)} = 1-\int_{-\infty}^{\infty}\phi(z)\left(2\Phi\left(\frac{\rho}{\sqrt{2-\rho^2}}z\right)-1\right)dz,$$

which can be shown to be, for $0\leq\rho\leq 1$,

$$\Delta^{*(1)}_{(1,2)}(\rho) = 1-\frac{2}{\pi}\tan^{-1}\left(\frac{\rho}{\sqrt{2-\rho^2}}\right).$$

For example, $\Delta^{*(1)}_{(1,2)}(0)=1$, $\Delta^{*(1)}_{(1,2)}(0.4)=0.81745$, $\Delta^{*(1)}_{(1,2)}(0.5)=0.76995$, $\Delta^{*(1)}_{(1,2)}(0.8)=0.61722$ & $\Delta^{*(1)}_{(1,2)}(1)=0.5$.

Thus, the actual values of the $\Delta^{*(1)}_{(i,i+1)}$ can be approximated once the value of ρ is estimated. The approximate values are compared to the actual values of $\Delta^{(1)}_{(i,i+1)}$ given in the table.

In general, the value $\Delta^{*(k)}_{(i,i+1)}$ is a function of the parameters of the underlying distribution; thus, it can be estimated by estimating these parameters. To investigate how accurate is judgment ranking, the estimator of $\Delta^{*(k)}_{(i,i+1)}$ can be statistically compared to the actual known value of $\Delta^{(k)}_{(i,i+1)}$. For more details about estimating Δ, see Weitzman (1970), Reiser and Faraggi (1999), and Clemons and Bradley, (2000).

Alternatively, in view of the interpretations of Δ given in the introduction, that regarded Δ as the probability of misclassification. $\Delta_{(i,i+1)}^{*(k)}$ can be estimated based on the obtained data. For example, if $m=2$, then each obtained measurement can be classified as being in the lower or the upper stratum or in the "indifference" zone. The proportion of values that are classified to be in the "indifference zone" (i.e., the overlapping area) is an estimator of Δ^*. Some statistical technique can be then used to see if there is a significant difference between the values of Δ^* and Δ. For example, the level of yellowness can be judged as low or high based on the color of the face before making any measurements. With $m=2$ and $h=50$, $r=1$, we have 100 individuals 50 of them classified as having low yellowness level and 50 as having high yellowness level. Now, exact level of yellowness can be determined for each individual using some medical tests and the actual classification of each individual can be obtained. An estimate of Δ is the proportion of individuals with low yellowness level but classified wrongly as individuals with high yellowness level + the proportion of individuals with high yellowness level but classified wrongly as individuals with low yellowness level.

The above uses and other possible uses of Δ in either the design or the analysis stage are currently under consideration by the second author.

4. Concluding Remarks

The MSRSS is a very useful technique for stratifying the population. We noted that as the number of stages increases, the overlapping coefficient between the densities of consecutive data points decreases. Thus, eventually, MSRSS divides the population into almost non-overlapping strata. The value of the overlapping coefficient at stage r is derived based on the previous stage. Also, it is noted the overlapping coefficient is independent of the underlying distribution. The overlapping coefficient between adjacent judgment order statistics was shown to be larger than the overlapping between the corresponding exact order statistics; they depend on the underlying distribution. Possible uses of Δ at the planning (Design) and the Inference (Analysis) stage are addressed and put forward for possible future research.

Acknowledgments

Our sincere thanks are due to the referees and Associate Editor who provided many critical comments and helpful suggestions which significantly improved the original version of the article.

References

Al-Saleh, M. F. (2004). Steady state ranked set sampling and parametric estimation. *J. Statist Plann. Infer.* 123:83–95.

Al-Saleh, M. F. (2007). On the similarity structure of order statistics. *Commun. Statist. Theor. Meth.* 36:1433–1439.

Al-Saleh, M. F., Ahmed, A. (2007). Estimating of the means of the bivariate normal using moving extreme ranked set sampling with concomitant variable. *Statist. Pap.* 48:179–195.

Al-Saleh, M. F., Al-Kadiri, M. (2000). Double-ranked set sampling. *Statist. Probab. Lett.* 48:205–212.

Al-Saleh, M. F., Al-Omari, A. (2002). Multistage ranked set sampling. *J. Statist. Plann. Infer.* 102:273–286.

Al-Saleh, M. F., Diab, Y. (2009). Estimation of the parameters of Downton's bivariate exponential distribution using different RSS schemes. *J. Statist. Plann. Infer.* 139:277–286.

Al-Saleh, M. F., Samawi, H. (2007). Inclusion probability in ranked set sampling from finite population. *Test* 16:198–209.

Al-Saleh, M. F., Samawi, H. (2009). On estimating the odds using moving extreme ranked set sampling. *Statist. Methodol.* 7:133–140.

Al-Saleh, M. F., Samuh, M. H. (2008). On multistage ranked set sampling for distribution and median estimation. *Computat. Statist. Data Anal.* 52:2066–2078.

Al-Saleh, M. F., Zheng, G. (2002). Estimation of bivariate characteristics using ranked set sampling. *Austral. NZ J. Statist.* 44:221–232.

Chen, Z., Bai, Z., Sinha, B. K. (2003). *Ranked Set Sampling Theory and Applications*. New York: Springer.

Clemons, T. E., Bradley, E. L., Jr. (2000). A nonparametric measure of the overlapping coefficient. *Comput. Statist. Data Anal.* 34:51–61.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets. *Austral. J. Agricult. Res.* 3:385–390.

McIntyre, G. A. (2005). A method for unbiased selective sampling using ranked sets. *Amer. Statistician* 59:230–232.

Oztork, O. (2010). Nonparametric maximum likelihood estimation of within-set ranking errors in ranked set sampling. *J. Nonparametric Statist.* 11:823–840.

Reiser, B., Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: the normal variance case. *J. Roy. Stat. Soc. Ser. (The Statistician)* 48:413–418.

Samawi, H., Al-Saleh, M. F., Al-Saidy, O. (2009). The matched pairs sign test using bivariate ranked set sampling for different ranking based schemes. *Statist. Methodol.* 6:397–407.

Takahasi, K., Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Ann. Instit. Statist. Math.* 20:1–31.

Weitzman, M. S. (1970). Measure of the overlap of income distribution of white and Negro families in the United States. *Technical Report No. 22*, U.S. Department of Commerce, Bureau of the Census, Washington, D.C.

Yang, S. S. (1977). General distribution theory of the concomitants of order statistics. *Ann. Statist.* 5:996–1002.

Zheng, G., Al-Saleh, M. F. (2002). Modified maximum likelihood estimator based on ranked set sampling. *Ann. Instit. Statist. Math.* 54:641–658.