Communications for Statistical Applications and Methods
2016, Vol. 23, No. 2, 119–129

http://dx.doi.org/10.5351/CSAM.2016.23.2.119
Print ISSN 2287-7843 / Online ISSN 2383-4757

Nonparametric confidence intervals for quantiles based on a modified ranked set sampling

Hakime Morabbi[a], Mostafa Razmkhah[1,a], Jafar Ahmadi[a]

[a]Department of Statistics, Ferdowsi University of Mashhad, Iran

Abstract

A new sampling method is introduced based on the idea of a ranked set sampling scheme in which taken samples in each set are dependent on previous ones. Some theoretical results are presented and distribution-free confidence intervals are derived for the quantiles of any continuous population. It is shown numerically that the proposed sampling scheme may lead to 95% confidence intervals (especially for extreme quantiles) that cannot be found based on the ordinary ranked set sampling scheme presented by Chen (2000) and Balakrishnan and Li (2006). Optimality aspects of this scheme are investigated for both coverage probability and minimum expected length criteria. A real data set is also used to illustrate the proposed procedure. Conclusions are eventually stated.

Keywords: order statistics, ranked set sampling, truncated distribution, optimality, extreme quantiles, coverage probability, minimum expected length, distribution–free confidence interval

1. Introduction

Ranked set sampling (RSS) is a cost efficient technique for data collection when sampling units can be ranked easily without actual measurement. RSS methodology was introduced by McIntyre (1952) to estimate the population mean. In this scheme, k independent sets each contains k units are randomly selected from the population. Next, the units of each set are visually sorted in ascending order and only the j^{th} ($j = 1, 2, \ldots, k$) order statistic of the j^{th} set is measured. When judgment ranking is accurate, the selected units are actually a sample of k independent order statistics. This process can be replicated m times to yield a sample of size mk.

A more detailed mathematical development was presented by Takahasi and Wakimoto (1968). Since then, there have been numerous parametric and nonparametric inferential procedures based on RSS in the literature. Barnett and Moore (1997), Barnett (1999), Chen *et al.* (2005, 2006), Chen *et al.* (2004), and Wolfe (2004) may be seen as landmark contributions. Recently, RSS was developed and modified by many authors in order to improve the estimation of population parameters. Samawi *et al.* (1996) and Samawi and Muttlak (1996) used extreme RSS method to estimate population mean and ratio, respectively. Al-Saleh and Al-Omari (2002) presented multistage stratified RSS to increase the efficiency of estimating the population mean for the specific value of sample size. Kadilar *et al.* (2009) obtained a ratio estimator for the population mean using RSS and showed that it is more efficient than the corresponding estimator based on simple random sampling. Al-Saleh and Samawi (2010) discussed estimating the odds using moving extreme RSS. Jafari Jozani *et al.* (2012) studied the problem of reducing the bias of the ratio estimator for the population mean in the RSS.

[1] Corresponding author: Department of Statistics, Ferdowsi University of Mashhad, P. O. Box 1159, Mashhad 91775, Iran. E-mail: razmkhahm@um.ac.ir

Published 31 March 2016 / journal homepage: http://csam.or.kr

The majority of RSS research has been concentrated on population mean estimation; however, other population parameters have been received less attention. We are interested in estimating population quantiles that are important in many fields such as environmental studies, quality control, and industrial destructive testing. Let $F(\cdot)$ denote the cumulative distribution function (cdf) of the baseline population, then, the quantile of order p is defined by $\xi_p = \inf\{x : F(x) \geq p\}$. Some authors have studied the quantiles based on RSS. Chen (2000) found that RSS method can substantially improve the efficiency of the quantile estimators rather than simple random sampling. Ozturk and Deshpande (2006) provided nonparametric exact confidence intervals (CIs) for quantiles based on RSS with higher coverage probability as well as a shorter expected length than simple random sampling. The CIs for quantiles and tolerance intervals based on ordered ranked set data were discussed by Balakrishnan and Li (2006). Mahdizadeh and Arghami (2012) used RSS method to introduce a mean-corrected quantile estimator.

In an ordinary RSS, k sets of simple random samples for each of size k are taken from the underlying distribution where each set is independent from the previous one. In this paper, we propose a new sampling method that use the idea of RSS the methodology so that each set depends on the previous one. We call this *dependent RSS* (DRSS). In this scheme, one may first take an initial sample of size n from an infinite population visually and arrange the units in ascending order with judgment. Then, depend on the main goal of the inference, certain unit from the initial sample is measured and accordingly the second sample is taken. If the estimation of the upper quantiles is of interest, it seems reasonable to measure an upper order statistic of the initial sample of size n, denoted its value by $x^{(1)}$. Then, for the second sample, choose n other elements of the underlying population such that their judged values are greater than $x^{(1)}$. In this sample, the same ranked order statistic as recorded in the initial sample is measured as new data point. Similarly, the third judging sample is taken using the information obtained from the second sample. This process is continued until the predetermined number of observations are measured. We call this procedure and *upper DRSS* (UDRSS) scheme. Logically, the observations less than the previous ones are collected if the estimation of the lower quantiles is of interest. Similarly, we call this sampling scheme as *lower DRSS* (LDRSS).

Suppose we would like to measure the amount of plutonium found in soil for agricultural studies. In such experiments, it is obvious that the high value of plutonium found in soil tends to decrease the amount of products. This indicates that we need to pay attention to land which the plutonium extent is lower than a certain value; consequently, it is beneficial to concentrate on extreme upper quantiles. It also must be noted that a radiochemical analysis of soil samples for actual measurements on plutonium in soil is expensive as well as hazardous. But the samples can be ranked according to value of a surrogate variable, viz., the field instrument for the determination of low energy radiation is counted per minute where the soil samples are taken for radiochemical analysis (Deshpande, 2013). Therefore, one may construct distribution-free CIs for upper quantiles using the UDRSS method. Chen (2000) and Balakrishnan and Li (2006) found, respectively, approximate and distribution-free CIs for quantiles based on ordinary RSS. Tables 1 and 2 based on one-cycle and Tables 7 and 8 based on two-cycle ordinary RSS, given by Balakrishnan and Li (2006), do not contain 90% and 95% CIs for extreme quantiles, respectively. We show that the proposed sampling scheme in this paper resolves this gap. Indeed, one can construct 95% CIs for extreme upper and lower quantiles of the underlying population using UDRSS and LDRSS schemes, respectively.

The rest of this paper is organized as follows. In Section 2, the model of interest is explained and also the fundamental results are presented. Distribution-free CIs for upper and lower quantiles are constructed in Section 3. The numerical results are presented in Section 4 and some optimality aspects in the coverage probability and minimum expected length of the proposed sampling scheme

are investigated. Section 5 illustrates the proposed procedure in the paper through a real data set. Finally, some conclusions and topics of future study are stated.

2. Model description

Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed (iid) random variables from an absolutely continuous cdf $F(\cdot)$ with the corresponding probability density function (pdf) $f(\cdot)$. To attain an UDRSS [or LDRSS] data set, the following steps must be performed:

1. For given n, the random variables $X_1, \ldots, X_n$ are visually ranked in ascending order.
2. For given j_2 [or j_1], the j_2^{th} [or j_1^{th}] ranked observation, denoted by $X_{j_2:n}^{(1)}$ [or $X_{j_1:n}^{(1)}$] is measured exactly.
3. For given $X_{j_2:n}^{(1)} = x_{j_2:n}^{(1)}$ [or $X_{j_1:n}^{(1)} = x_{j_1:n}^{(1)}$] in Step 2, a new random sample of size n is independently taken from a subpopulation with the cdf

$$1 - \frac{\bar{F}(y)}{\bar{F}\left(x_{j_2:n}^{(1)}\right)}, \quad y > x_{j_2:n}^{(1)} \left[\text{or } \frac{F(y)}{F\left(x_{j_1:n}^{(1)}\right)}, \; y < x_{j_1:n}^{(1)}\right],$$

   when the UDRSS [or LDRSS] scheme is used. To obtain the second sample, some elements of the underlying population are sequentially considered and those with the judged values greater than $x_{j_2:n}^{(1)}$ [or less than $x_{j_1:n}^{(1)}$] are taken.
4. Without yet knowing any values for the variable of interest, rank visually the units of the sample taken in Step 3 and measure the exact value of the j_2^{th} [or the j_1^{th}] order statistic, which is denoted by $X_{j_2:n}^{(2)}$ [or $X_{j_1:n}^{(2)}$].
5. This sampling procedure continues to obtain sequentially k dependent data points denoted by $\{X_{j_2:n}^{(1)}, X_{j_2:n}^{(2)}, \ldots, X_{j_2:n}^{(k)}\}$ [or $\{X_{j_1:n}^{(1)}, X_{j_1:n}^{(2)}, \ldots, X_{j_1:n}^{(k)}\}$], where k is a pre-specified integer number.

Here, we emphasize that unlike ordinary RSS, $X_{j_2:n}^{(r)}$s [or $X_{j_1:n}^{(r)}$s], $r = 1, \ldots, k$, are dependent and ordered in r, i.e., $X_{j_2:n}^{(r)} \leq X_{j_2:n}^{(r+1)}$ [or $X_{j_1:n}^{(r+1)} \leq X_{j_1:n}^{(r)}$], $r = 1, \ldots, k-1$, with probability one. Note that the success of this sampling scheme depends on the accuracy of ranking the units in each set. For this reason, the set size n must be kept small to reduce errors in judgement ranking. It is obvious that a small number of sampling units can be easily ordered with respect to the variable of interest. Nevertheless, Balakrishnan and Li (2006) obtained 95% CIs for $\xi_{0.2}$ an $\xi_{0.8}$, when $n = 9$ based on one-cycle ordinary RSS, which is rather large sample size. Even by applying two cycles of an ordinary RSS, they could not derive 95% CIs, when $n \leq 5$.

We now investigate the cdf and pdf of the observed order statistic at the r^{th} $(1 \leq r \leq k)$ stage of the proposed sampling schemes. At the r^{th} stage of an UDRSS, suppose that $X_1^{(r)}, \ldots, X_n^{(r)}$ $(1 \leq r \leq k)$ represent a simple random sample of size n from a subpopulation of the cdf $F(\cdot)$ truncated at the left of $x_{j_2:n}^{(r-1)}$, the observed value of $X_{j_2:n}^{(r-1)}$, with $X_{j_2:n}^{(0)} = F^{-1}(0)$. Therefore, $X_{j_2:n}^{(1)}$ stands for the j_2^{th} order statistic in a random sample of size n from the cdf $F(\cdot)$. Moreover, for $2 \leq r \leq k$, given $X_{j_2:n}^{(r-1)} = x_{j_2:n}^{(r-1)}$ $(1 \leq j_2 \leq n)$, the corresponding conditional cdf and pdf of the subpopulation of the cdf $F(\cdot)$ at the r^{th} stage, for $x > x_{j_2:n}^{(r-1)}$, are given as

$$G\left(x|x_{j_2:n}^{(r-1)}\right) = 1 - \frac{\bar{F}(x)}{\bar{F}\left(x_{j_2:n}^{(r-1)}\right)}$$

and

$$g\left(x|x_{j_2:n}^{(r-1)}\right) = \frac{f(x)}{\bar{F}\left(x_{j_2:n}^{(r-1)}\right)},$$

respectively, where $\bar{F}(\cdot) = 1 - F(\cdot)$ denotes the survival function of X. So, using the conditional argument, the cdf of $X_{j_2:n}^{(r)}$, for $2 \le r \le k$, can be written as

$$\begin{aligned} G_{j_2:n}^{(r)}(x) &= P\left(X_{j_2:n}^{(r)} \le x\right) \\ &= \sum_{t=j_2}^{n} \int_{-\infty}^{+\infty} P\left(A_t^{(r)}(x) \middle| X_{j_2:n}^{(r-1)} = y\right) g_{j_2:n}^{(r-1)}(y)dy, \end{aligned}$$

where $A_t^{(r)}(x)$ means the event that exactly t of $X_1^{(r)}, \ldots, X_n^{(r)}$ are at most x. Therefore,

$$G_{j_2:n}^{(r)}(x) = \sum_{t=j_2}^{n} \binom{n}{t} \int_{-\infty}^{x} [G(x|y)]^t [1 - G(x|y)]^{n-t} g_{j_2:n}^{(r-1)}(y)dy. \tag{2.1}$$

Also, by differentiating from both sides of (2.1) with respect to x and doing some algebraic calculations, the pdf of $X_{j_2:n}^{(r)}$ is given by

$$g_{j_2:n}^{(r)}(x) = j_2 \binom{n}{j_2} \int_{-\infty}^{x} g(x|y)[G(x|y)]^{j_2-1}[1 - G(x|y)]^{n-j_2} g_{j_2:n}^{(r-1)}(y)dy.$$

For an LDRSS, it can also be shown that for given j_1, the marginal pdf and cdf of $X_{j_1:n}^{(r)}$ $(2 \le r \le k)$, the j_1^{th} order statistic in a random sample of size n from a subpopulation of the cdf $F(\cdot)$ truncated at the right of $x_{j_1:n}^{(r-1)}$, the observed value of $X_{j_1:n}^{(r-1)}$, are given by

$$h_{j_1:n}^{(r)}(x) = j_1 \binom{n}{j_1} \int_{x}^{+\infty} h(x|y)[H(x|y)]^{j_1-1}[1 - H(x|y)]^{n-j_1} h_{j_1:n}^{(r-1)}(y)dy,$$

and

$$H_{j_1:n}^{(r)}(x) = H_{j_1:n}^{(r-1)}(x) + \sum_{t=j_1}^{n} \binom{n}{t} \int_{x}^{+\infty} [H(x|y)]^t [1 - H(x|y)]^{n-t} h_{j_1:n}^{(r-1)}(y)dy, \tag{2.2}$$

respectively. Here, for $x < y$, $H(x|y) = F(x)/F(y)$ and $h(x|y) = f(x)/F(y)$, respectively, stand for the conditional cdf and pdf of the subpopulation of the cdf F at the r^{th} stage, given $X_{j_1:n}^{(r-1)} = y$ $(1 \le j_1 \le n)$.

Notice that the UDRSS and LDRSS can be replicated m times independently to get data sets of size $N = mk$, which lead to

$$\mathbf{X}^{UDRSS} = \left\{ \underbrace{X_{j_2:n}^{(1,1)}, \ldots, X_{j_2:n}^{(k,1)}}, \underbrace{X_{j_2:n}^{(1,2)}, \ldots, X_{j_2:n}^{(k,2)}}, \ldots, \underbrace{X_{j_2:n}^{(1,m)}, \ldots, X_{j_2:n}^{(k,m)}} \right\} \tag{2.3}$$

and

$$\mathbf{X}^{LDRSS} = \left\{ \underbrace{X_{j_1:n}^{(1,1)}, \ldots, X_{j_1:n}^{(k,1)}}, \underbrace{X_{j_1:n}^{(1,2)}, \ldots, X_{j_1:n}^{(k,2)}}, \ldots, \underbrace{X_{j_1:n}^{(1,m)}, \ldots, X_{j_1:n}^{(k,m)}} \right\}, \tag{2.4}$$

respectively, where $X^{(r,s)}_{j_2:n}$ [or $X^{(r,s)}_{j_1:n}$] represents the j_2^{th} [or j_1^{th}] order statistic in a random sample of size n from the r^{th} $(1 \le r \le k)$ stage of an UDRSS [or LDRSS] at the s^{th} $(1 \le s \le m)$ cycle. Indeed, the data sets in (2.3) and (2.4) therefore contain the UDRSS and LDRSS schemes with k dependent sets which are performed in m independent cycles. The goal of this paper is to construct CIs for upper and lower population quantiles; therefore, we arrange the elements of $\mathbf{X}^{UDRSS}$ and $\mathbf{X}^{LDRSS}$ in ascending order and use the induced order statistics as confidence bounds. To derive the corresponding marginal cdfs, the Theorem 3 of Chahkandi *et al.* (2014) has been simplified for a special case as stated in the following lemma.

Lemma 1. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be m iid continuous random vectors for which $\mathbf{Y}_s = (Y_{s,1}, \ldots, Y_{s,k})$ such that $Y_{s,1} < \cdots < Y_{s,k}$ with probability one, for $1 \le s \le m$ and $k \ge 2$. Then, the marginal cdf of $Y_{i:mk}$, the i^{th} order statistic among the data set $\{Y_{s,r}, 1 \le s \le m, 1 \le r \le k\}$, is given by*

$$\begin{aligned} P(Y_{i:mk} \le x) &= \sum_{t=i}^{mk} \sum_{h_0=\alpha_0(t)}^{\alpha'_0(t)} \cdots \sum_{h_{k-2}=\alpha_{k-2}(t)}^{\alpha'_{k-2}(t)} m! \prod_{s=0}^{k} \frac{1}{h_s!} \left(\psi^{(k-s)}(x) - \psi^{(k-s+1)}(x)\right)^{h_s} \\ &= \Lambda_i \left(\psi^{(0)}(x), \ldots, \psi^{(k+1)}(x)\right), \ \textit{say}, \end{aligned} \tag{2.5}$$

where $\psi^{(0)}(x) = 1$, $\psi^{(k+1)}(x) = 0$ and for $1 \le r \le k$, we have

$$\psi^{(r)}(x) = P(Y_{1,r} \le x). \tag{2.6}$$

Moreover, for $0 \le i \le k-2$,

$$\alpha_i(t) = \max\left\{0, t - \sum_{l=0}^{i-1}(i+1-l)h_l - m(k-i-1)\right\} \quad \textit{and} \quad \alpha'_i(t) = \left[\frac{t - \sum_{l=0}^{i-1}(k-l)h_l}{k-i}\right],$$

where $[u]$ stands for the integer part of u. Furthermore,

$$h_{k-1} = t - \sum_{l=0}^{k-2}(k-l)h_l \quad \textit{and} \quad h_k = m - t + \sum_{l=0}^{k-2}(k-1-l)h_l.$$

3. Confidence intervals for quantiles

In this section, the CIs for upper and lower quantiles of the underlying population are obtained basis on the UDRSS and LDRSS schemes, respectively. All results are shown to be distribution-free.

The data set $\mathbf{X}^{UDRSS}$ S in (2.3) are preferred for the construction of the CIs for upper quantiles of cdf F. Let us denote the i^{th} order statistic of this data set by $X^{UDRSS}_{i:mk}$. Note that for the s^{th} cycle $(1 \le s \le m)$, we have $X^{(1,s)}_{j_2:n} < X^{(2,s)}_{j_2:n} < \cdots < X^{(k,s)}_{j_2:n}$, with probability one. Therefore, by Lemma 1 and because for $1 \le r \le k$, $Y_{1,r}$ in (2.6) is equivalent to $X^{(r,1)}_{j_2:n}$ in $\mathbf{X}^{UDRSS}$, it can be deduced that for fixed value of $k \ge 2$ and $m \ge 1$, we have

$$P\left(X^{UDRSS}_{i:mk} \le \xi_p\right) = \Lambda_i \left(\psi^{(0)}_u(j_2; p), \ldots, \psi^{(k+1)}_u(j_2; p)\right), \tag{3.1}$$

where $\Lambda_i(\cdot)$ is as defined in (2.5) and

$$\psi^{(r)}_u(j_2; p) = \begin{cases} 1, & r = 0, \\ F_{j_2:n}(\xi_p), & r = 1, \\ G^{(r)}_{j_2:n}(\xi_p), & 2 \le r \le k, \\ 0, & r = k+1, \end{cases}$$

where $F_{j:n}(\cdot)$ is the cdf of the j^{th} order statistic in a simple random sample of size n from the cdf $F(\cdot)$ and $G^{(r)}_{j_2:n}(\cdot)$ is as defined in (2.1). Since $\xi_p = F^{-1}(p)$, it is clear that

$$F_{j_2:n}(\xi_p) = \sum_{t=j_2}^{n} \binom{n}{t} p^t (1-p)^{n-t}. \tag{3.2}$$

Moreover, using (2.1) by doing some algebraic calculations, it can be shown that

$$\begin{aligned} G^{(r)}_{j_2:n}(\xi_p) = \sum_{t=j_2}^{n} \binom{n}{t} \left\{ j_2 \binom{n}{j_2} \right\}^{r-1} \int_0^p \int_0^{\nu_1} \cdots \int_0^{\nu_{r-2}} \left(1 - \frac{1-p}{1-\nu_1}\right)^t \left(\frac{1-p}{1-\nu_1}\right)^{n-t} \\ \times \left\{ \prod_{i=2}^{r} \frac{1}{1-\nu_i} \left(1 - \frac{1-\nu_{i-1}}{1-\nu_i}\right)^{j_2-1} \left(\frac{1-\nu_{i-1}}{1-\nu_i}\right)^{n-j_2} \right\} d\nu_{r-1} \cdots d\nu_1, \end{aligned}$$

such that $\nu_r = 0$. Consequently, $\psi^{(r)}_u(j_2; p)$ depends only on n, j_2 and p and is free of the baseline cdf $F(\cdot)$. Therefore, the random interval $(X^{UDRSS}_{i_1:mk}, X^{UDRSS}_{i_2:mk})$ can be considered as a distribution-free CI for ξ_p with coverage probability

$$\gamma_u(i_1, i_2; j_2; p) = \Lambda_{i_1}\left(\psi^{(0)}_u(j_2; p), \ldots, \psi^{(k+1)}_u(j_2; p)\right) - \Lambda_{i_2}\left(\psi^{(0)}_u(j_2; p), \ldots, \psi^{(k+1)}_u(j_2; p)\right), \tag{3.3}$$

which depends on the values of i_1, i_2, n, k, j_2 and p, not on the baseline cdf $F(\cdot)$.

The data set $\mathbf{X}^{LDRSS}$ used to construct appropriate CIs for lower quantiles in (2.4) can also be. Consider an experiment about assessing spray reserves on the leaves of trees presented by Murray *et al.* (2000). After a certain time of completion spraying, a chemical experiment will determine the remaining pesticide on leaves. The leaves will be more susceptible to pests if it is lower than a certain value. To attain the results of this experiment, a large number of leaves must be harvested for the measurement of remaining pesticide on leaves. It is obvious that studies of this type involve an extensive sampling effort. However, visual ranking is possible because leaf deposits can be viewed subsequently under ultraviolet light if trees are sprayed with a fluorescent tracer. So, the percentage of leaves with lower remaining pesticide is of interest. In such situations, it is valuable to obtain distribution-free CIs for extreme lower quantiles by implementing a LDRSS scheme.

Now, denote the i^{th} order statistic of the data set in (2.4) by $X^{LDRSS}_{i:mk}$. Here, for the s^{th} $(1 \le s \le m)$ cycle, we get $X^{(1,s)}_{j_1:n} > \cdots > X^{(k,s)}_{j_1:n}$, with probability one. Therefore, by use of (2.5) and the fact that for $1 \le r \le k$, $Y_{1,r}$ in (2.6) is equivalent to $X^{(k-r+1,1)}_{j_1:n}$ in $\mathbf{X}^{LDRSS}$, it can be shown that the interval $(X^{LDRSS}_{i_1:mk}, X^{LDRSS}_{i_2:mk})$ is a CI for ξ_p with the following coverage probability

$$\gamma_l(i_1, i_2; j_1; p) = \Lambda_{i_1}\left(\psi^{(k+1)}_l(j_1; p), \ldots, \psi^{(0)}_l(j_1; p)\right) - \Lambda_{i_2}\left(\psi^{(k+1)}_l(j_1; p), \ldots, \psi^{(0)}_l(j_1; p)\right), \tag{3.4}$$

where

$$\psi^{(r)}_l(j_1; p) = \begin{cases} 0, & r = 0, \\ F_{j_1:n}(\xi_p), & r = 1, \\ H^{(r)}_{j_1:n}(\xi_p), & 2 \le r \le k, \\ 1, & r = k+1, \end{cases}$$

such that $F_{j_1:n}(\cdot)$ is as defined in (3.2). Furthermore, using (2.2) by performing some algebraic calculations, we get

$$H^{(r)}_{j_1:n}(\xi_p) = F_{j_1:n}(\xi_p) + \sum_{s=2}^{r}\sum_{t=j_1}^{n}\binom{n}{t}\left\{j_1\binom{n}{j_1}\right\}^{s-1}\int_p^1\int_{\nu_1}^1\cdots\int_{\nu_{s-2}}^1\left(\frac{p}{\nu_1}\right)^t\left(1-\frac{p}{\nu_1}\right)^{n-t}$$
$$\times\left\{\prod_{i=2}^{s}\frac{1}{\nu_i}\left(\frac{\nu_{i-1}}{\nu_i}\right)^{j_1-1}\left(1-\frac{1-\nu_{i-1}}{\nu_i}\right)^{n-j_1}\right\}d\nu_{s-1}\cdots d\nu_1,$$

such that $\nu_s = 1$. It is clear that the probability $\gamma_l(i_1, i_2; j_1; p)$ in (3.4) depends only on i_1, i_2, n, k, j_1 and p, not on the baseline cdf $F(\cdot)$. This means that the interval $(X^{LDRSS}_{i_1:mk}, X^{LDRSS}_{i_2:mk})$ is indeed a distribution-free CI for ξ_p.

4. Optimal confidence intervals

Here, we investigate the optimal indices for the proposed CIs in the sense of both coverage probability and minimum expected length. Toward this end, for given values of n, mk, p and α_0, the optimal CIs for upper quantiles based on the UDRSS scheme are determined such that the two-sided CI $(X^{UDRSS}_{i_1:mk}, X^{UDRSS}_{i_2:mk})$ has the coverage probability greater than $1-\alpha_0$, i.e.,

$$\gamma_u(i_1, i_2, j_2; p) \geq 1-\alpha_0. \tag{4.1}$$

We now choose as small as possible k and j_2 because the smallest number of experimental units leads to the lowest test cost. Therefore, the optimal values of (k, j_2, i_1, i_2) are determined through the following steps:

1. Set $k = 1$ and $j_2 = 1$.
2. Determine all pairs of (i_1, i_2), $1 \leq i_1 < i_2 \leq mk$, such that (4.1) holds.
3. Select the pair (i_1, i_2) that minimizes the difference $i_2 - i_1$. If there exist more than one pair of (i_1, i_2) with the same difference, then choose the one that minimizes the expected width $E(U_{i_2:mk} - U_{i_1:mk})$, where $U_{i:mk}$ is the i^{th} order statistic in an UDRSS data set of size mk from the uniform distribution on $(0, 1)$. The final selected value for (i_1, i_2) is optimal, which is denoted by (i^*_1, i^*_2).
4. If there is no any pair of (i_1, i_2), then, while $j_2 < n$, put $j_2 = j_2 + 1$ and go to the Step 2.
5. If there is no any optimal value for j_2, then put $k = k + 1$ and go to the Step 2.
6. The optimal values of k and j_2 are denoted by k^* and j^*_2, respectively.

Similarly, the optimal values of (k, j_1, i_1, i_2), denoted by $(k^*, j^*_1, i^*_1, i^*_2)$, may be derived to obtain the CIs for lower quantiles based on the LDRSS scheme such that

$$\gamma_l(i_1, i_2, j_1; p) \geq 1-\alpha_0. \tag{4.2}$$

In this case, an analogous algorithm as mentioned above should be performed, with this differences that at first we set $j_1 = n$ in the Step 1 and looking for the pairs of (i_1, i_2) in Step 2 such that (4.2) holds. Then gradually decrease j_1 until $j_1 \geq 1$, i.e., in Step 4, we put $j_1 = j_1 - 1$. This strategy shows that the smallest number of experimental units is needed for the LDRSS scheme.

Table 1: Values of $(k^*, j_2^*, i_1^*, i_2^*)$ to obtain 95% CIs for ξ_p $(p \geq 0.5)$ based on UDRSS, for $mk = 12$

n		p						
		0.50	0.70	0.80	0.85	0.90	0.95	0.99
3	k^*	1	1	1	1	1	2	3
	j_2^*	2	3	3	3	3	3	3
	(i_1^*, i_2^*)	(3, 10)	(1, 8)	(3, 10)	(5, 12)	(6, 12)	(5, 12)	(5, 12)
4	k^*	1	1	1	1	1	2	3
	j_2^*	2 or 3	3 or 4	4	4	4	4	4
	(i_1^*, i_2^*)	(1, 9) or (1, 8)	(4, 11) or (1, 7)	(2, 9)	(3, 10)	(5, 12)	(4, 11)	(5, 12)
5	k^*	1	1	1	1	1	1	2
	j_2^*	3	4	4 or 5	5	5	5	5
	(i_1^*, i_2^*)	(3, 10)	(3, 10)	(6, 12) or (1, 8)	(2, 9)	(4, 11)	(5, 12)	(6, 12)

CI = confidence interval, UDRSS = upper dependent ranked set sampling.

Table 2: Values of $(k^*, j_1^*, i_1^*, i_2^*)$ to obtain 95% CI for ξ_p $(p \leq 0.5)$ based on LDRSS, for $mk = 12$

n		p						
		0.01	0.05	0.10	0.15	0.20	0.30	0.50
3	k^*	3	2	1	1	1	1	1
	j_1^*	1	1	1	1	1	1	2
	(i_1^*, i_2^*)	(1,8)	(1,8)	(1,7)	(1,8)	(3,10)	(3,11)	(3,10)
4	k^*	3	2	1	1	1	1	1
	j_1^*	1	1	1	1	1	1 or 2	2 or 3
	(i_1^*, i_2^*)	(1,8)	(3,10)	(1,8)	(3,10)	(4,11)	(6,12) or (3,9)	(5,12) or (1,8)
5	k^*	2	1	1	1	1	1	1
	j_1^*	1	1	1	1	1 or 2	2	3
	(i_1^*, i_2^*)	(1,7)	(1,8)	(2,9)	(2,10)	(5,12) or (1,7)	(3,10)	(3,10)

CI = confidence interval, LDRSS = lower dependent ranked set sampling.

The optimal four-tuples $(k^*, j_2^*, i_1^*, i_2^*)$ and $(k^*, j_1^*, i_1^*, i_2^*)$ may be computed using the above algorithm to obtain 95% CIs for ξ_p based on the UDRSS and LDRSS schemes, respectively, for given values of n, mk and p. Tables 1 and 2 presents the results when $mk = 12$.

From Tables 1 and 2, the following results are deduced:

- The 95% CIs for ξ_p $(0.1 < p < 0.9)$ may obtain, when $k^* = 1$, whereas for $p \geq 0.95$ or $p \leq 0.05$, k^* should be greater than one. That is, an UDRSS or LDRSS should be performed when the 95% CIs may not be obtained using the ordinary RSS.

- By comparing the entries of these tables, a symmetric property is seen. Indeed, for given n, the values of k^* to estimate ξ_p and ξ_{1-p} are the same. Moreover, the value of j_2^* to estimate ξ_p is equal to $(n - j_1^* + 1)$ to estimate ξ_{1-p}.

The results of Tables 1 and 2 have been obtained based on 12 data points. The optimal values have also been computed for data sets with 6 and 24 data points and similar results derived; however, we have not presented the details to summarize the results. All results in this section, have been computed by using R.3.1.2.

Remark 1. Denoting the r^{th} $(1 \leq r \leq k)$ observation at the s^{th} $(1 \leq s \leq m)$ cycle of UDRSS and LDRSS data sets from uniform distribution by $U_{j_2:n}^{(r,s)}$ and $L_{n-j_2+1:n}^{(r,s)}$, respectively, we have

$$U_{j_2:n}^{(r,s)} \stackrel{d}{=} 1 - L_{n-j_2+1:n}^{(r,s)},$$

where $\stackrel{d}{=}$ stands for identical in distribution. This confirms the symmetric property in the entries of Tables 1 and 2.

5. Application on a real data set

To illustrate the proposed procedure in the paper, we use a real data set corresponding to the maximum weekend car speeds presented at (Castillo *et al.*, 2005, p.17). These data are as follows:

68.9, 81.7, 67.7, 109.3, 118.9, 127.7, 109.4, 128, 102.9, 114.3, 102.3, 65.1, 120.4, 83.9, 83, 90.8, 120.9, 93.2, 81.1, 134.3, 93.8, 84.4, 117, 88, 98.9, 93.4, 91.9, 85.8, 107.1, 105.7, 84.8, 90.4, 90.1, 101.3, 90.4, 104.5, 92.4, 85.8, 93.3, 101.2, 88.2, 98.4, 124.4, 165.8, 84.9, 88.2, 115, 102.6, 75, 141.3, 115.5, 96.6, 86.7, 81.5, 83.8, 91.5, 107.7, 88.7, 81.5, 119, 142.2, 82, 95.1, 93, 92.1, 99.9, 109.2, 81.3, 78.9, 116.2, 79.6, 97.1, 108.6, 83.6, 114.2, 139.4, 68.5, 103.4, 93, 93.8, 108.2, 88.6, 124.8, 85.6, 75.1, 80.9, 98, 88.4, 71.8, 164.5, 72.2, 100.9, 83.1, 83.5, 94.9, 89.3, 97.3, 90.7, 89.4, 88.1, 107, 120.2, 112, 108, 88.1, 101.7, 138.2, 118.8, 81, 110, 95.4, 100.7, 90.4, 99.5, 93.8, 101.4, 72.8, 71, 127.3, 73.9, 97.7, 87.5, 83.2, 122.9, 106, 118.7, 97.1, 129.2, 105.6, 132.7, 80.5, 79.6, 126.6, 92.6, 95.6, 104.1, 87.3, 111.1, 114.6, 91.4, 81.8, 98.9, 98.3, 106.8, 113.3, 97.1, 92.4, 121.3, 99, 71.7, 72.4, 85.6, 81.1, 95.2, 103.5, 87.4, 8.6, 91.9, 106.7, 130.8, 152.3, 83.4, 114.1, 93.9, 117, 108.5, 108.5, 81.2, 111.8, 123.6, 121.3, 76.2, 83.8, 91.4, 98.2, 90.6, 94.1, 84, 138.3, 98.2, 90.6, 64.1, 84, 138.3, 98.2, 99.1, 111.5, 104.5, 90.1, 78.5, 83.6, 98.5, 133.8, 80.4, 85.3, 115.1, 78.2, 102.3, 127.5, 79.5, 87.4, 74.3, 102.2, 75.3, 131.3.

To find distribution-free 95% CI for $\xi_{0.95}$, we use an UDRSS scheme with $n = 3$ and $mk = 12$. From Table 1, it is observed that the optimal values of j_2 and k are 3 and 2, respectively. Consequently, the appropriate value of m is 6. Therefore, to obtain the required UDRSS data set, the following algorithm is used:

1. The first three iid observations, i.e., 68.9, 81.7 and 67.7 are taken as the initial sample (first set) and since $j_2^* = n = 3$, their maximum is considered as the first data point, that is $X_{3:3}^{(1,1)} = 81.7$.

2. From a subpopulation of the main distribution in which the data are visually greater than $X_{3:3}^{(1,1)} = 81.7$, the second data point is picked out. Therefore, from after the fourth observation, the first three ones which are greater than $X_{3:3}^{(1,1)} = 81.7$ in judgment are drawn as the second set and only their maximum is recorded. This new observation is $X_{3:3}^{(2,1)} = 127.7$.

3. The Steps 1 and 2 are repeated 6 times using the remaining data to attain 12 data points. Note that in different cycles different subpopulations may arise.

Thus, we get

$$X^{UDRSS} = \{81.7, 127.7, 128, 165.8, 115.5, 142.2, 108.6, 139.4, 85.6, 164.5, 100.9, 120.2\}.$$

Again, using Table 1, the interval $(X_{5:12}^{UDRSS}, X_{12:12}^{UDRSS}) = (115.5, 165.8)$ is the optimal 95% CIs for $\xi_{0.95}$.

6. Conclusion

We introduced a new sampling scheme based on the methodology of the ordinary RSS to obtain a modified data set to construct CIs for extreme quantiles of any continuous population. However, the random samples in each set of an ordinary RSS are independent in the proposed sampling method,

each set was dependent on the previous one according to the main goal of the inference. In fact, we introduced the UDRSS and LDRSS schemes with k dependent samples, where the corresponding order statistics estimated upper and lower quantiles, respectively. The numerical study showed that the proposed sampling schemes can be used in the situations in which the ordinary RSS may not provide 95% CIs for ξ_p, especially when $p \geq 0.95$ or $p \leq 0.05$. Moreover, the optimal CIs were obtained. The proposed procedure can be extended to other cases that may be considered as future research works:

- All results of this paper were distribution-free. The model can be used to perform parametric inference about various parameters of underlying population not only the quantiles.
- The proposed procedure is useful for inference about the quality control indices.
- The problem of predicting future extreme order statistics can be studied based on observed order statistics of an DRSS.

Acknowledgements

We express our sincere thanks to the referees for their constructive suggestions and useful comments on the original version which resulted in this improved version of the paper.

References

Al-Saleh MF and Al-Omari AI (2002). Multistage ranked set sampling, *Journal of Statistical Planning and Inference*, **102**, 273–286.

Al-Saleh MF and Samawi H (2010). On estimating the odds using moving extreme ranked set sampling, *Statistical Methodology*, **7**, 133–140.

Balakrishnan N and Li T (2006). Confidence intervals for quantiles and tolerance intervals based on ordered ranked set samples, *Annals of the Institute of Statistical Mathematics*, **58**, 757–777.

Barnett V (1999). Ranked set sample design for environmental investigations, *Environmental and Ecological Statistics*, **6**, 59–74.

Barnett V and Moore K (1997). Best linear unbiased estimates in ranked-set sampling with particular reference to imperfect ordering, *Journal of Applied Statistics*, **24**, 697–710.

Castillo E, Hadi AS, Balakrishnan N, and Sarabia JM (2005). *Extreme Value and Related Models with Applications in Engineering and Science*, John Wiley & Sons, New York.

Chahkandi M, Ahmadi J, and Baratpour S (2014). Non-parametric prediction intervals for the lifetime of coherent systems, *Statistical Papers*, **55**, 1019–1034.

Chen H, Stasny EA, and Wolfe DA (2005). Ranked set sampling for efficient estimation of a population proportion, *Statistics in Medicine*, **24**, 3319–3329.

Chen H, Stasny EA, and Wolfe DA (2006). Unbalanced ranked set sampling for estimating a population proportion, *Biometrics*, **62**, 150–158.

Chen Z (2000). The efficiency of ranked-set sampling relative to simple random sampling under multi-parameter families, *Statistica Sinica*, **10**, 247–263.

Chen Z, Bai Z, and Sinha B (2004). *Ranked Set Sampling: Theory and Application*, Springer, New York.

Deshpande JV (2013). Ranked set sampling for environmental studies, Retrieved March 1, 2016, from: http://www.samsi.info/sites/default/files/Deshpande_march2013.pdf

Jafari Jozani M, Majidi S, and Perron F (2012). Unbiased and almost unbiased ratio estimators of the population mean in ranked set sampling, *Statistical Papers*, **53**, 719–737.

Kadilar C, Unyazici Y, and Cingi H (2009). Ratio estimator for the population mean using ranked set sampling, *Statistical Papers*, **50**, 301–309.

Mahdizadeh M and Arghami NR (2012). Quantile estimation using ranked set samples from a population with known mean, *Communications in Statistics - Simulation and Computation*, **41**, 1872–1881.

McIntyre GA (1952). A method for unbiased selective sampling, using ranked sets, *Australian Journal of Agricultural Research*, **3**, 385–390.

Murray RA, Ridout MS, and Cross JV (2000). The use of ranked set sampling in spray deposit assessment, *Aspects of Applied Biology*, **57**, 141–146.

Ozturk O and Deshpande JV (2006). Ranked-set sample nonparametric quantile confidence intervals, *Journal of Statistical Planning and Inference*, **136**, 570–577.

Samawi HM, Ahmed MS, and Abu-Dayyeh W (1996). Estimating the population mean using extreme ranked set sampling, *Biometrical Journal*, **38**, 577–586.

Samawi HM and Muttlak HA (1996). Estimation of ratio using rank set sampling, *Biometrical Journal*, **38**, 753–764.

Takahasi K and Wakimoto K (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of the Institute of Statistical Mathematics*, **20**, 1–31.

Wolfe DA (2004). Ranked set sampling: an approach to more efficient data collection, *Statistical Science*, **19**, 636–643.

Received August 8, 2015; Revised February 1, 2016; Accepted February 3, 2016