Sankhyā : The Indian Journal of Statistics
2015, Volume 77-A, Part 2, pp. 408-432

Nonparametric Confidence Intervals for Quantiles with Randomized Nomination Sampling

Mohammad Nourmohammadi, Mohammad Jafari Jozani and Brad C. Johnson
University of Manitoba, Winnipeg, Canada

Abstract

Rank-based sampling methods have a wide range of applications in environmental and ecological studies as well as medical research and they have been shown to perform better than simple random sampling (SRS) for estimating several parameters in finite populations. In this paper, we obtain nonparametric confidence intervals for quantiles based on randomized nomination sampling (RNS) from continuous distributions. The proposed RNS confidence intervals provide higher coverage probabilities and their expected length, especially for lower and upper quantiles, can be substantially shorter than their counterparts under SRS design. We observe that a design parameter associated with the RNS design allows one to construct confidence intervals with the exact desired coverage probabilities for a wide range of population quantiles without the use of randomized procedures. Theoretical results are augmented with numerical evaluations and a case study based on a livestock data set. Recommendations for choosing the RNS design parameters are made to achieve shorter RNS confidence intervals than SRS design and these perform well even when ranking is imperfect.

AMS (2000) *subject classification.* Primary 62G15; Secondary 62D05.
Keywords and phrases. Confidence interval, infinite population, order statistics, nomination sampling, imperfect ranking.

1 Introduction

In many applications, an important concern is to construct confidence intervals for quantiles of the underlying population. For example, interval estimation of the upper or lower quantiles is important for the assessment of the status of hazard waste sites (Yu and Lam, 1997), monitoring water quality (Kvam, 2003), evaluating mercury contamination in fish (Murff and Sager, 2006), studying the adverse effects of tobacco smoke exposure of mothers during the pregnancy on health indices of infants (Burgette and Reinter, 2012) or more generally any risk factor or an exposure index in environmental, ecological, and medical studies. In many of such applications, the measurement of the variable of interest is costly, time consuming

or destructive, but a small number of sampling units can be fairly accurately ordered with respect to a variable of interest without actual measurements on them and at little cost. This ranking process can be used to obtain more representative samples form the underlying population and might also lead to increased precision, decreased sampling cost, or both. Randomized nomination sampling (RNS) is one such rank-based sampling designs introduced by Jafari Jozani and Johnson (2012), who provided design based estimation procedures for use in finite population sampling and showed some optimality results over simple random sampling (SRS) and strict maxima (minima) nomination sampling schemes.

In this paper, we consider the construction of nonparametric confidence intervals for quantiles under RNS design when applied to infinite populations. In general, to obtain an RNS sample of size m, one selects m subsamples (of random sizes) from the population. Within each subsample, the elements are ranked, without actual measurement, based on some easy to measure auxiliary attribute or on visual inspection. We then select either the maximum (with probability ζ) or minimum (with probability $1-\zeta$) for actual measurement. A more formal description of this process appears later. When the probability of selecting the maximum from each subsample is $\zeta = 1$, we obtain maxima nomination sampling, which was introduced by Willemain (1980). When $\zeta = 0$, we select the minimum of each subsample with probability 1, which results in minima nomination sampling (Wells and Tiwari, 1990). Various forms of nomination sampling have been applied to the estimation of distribution functions (Boyles and Samaniego, 1986; Tiwari, 1988; Kvam and Samaniego, 1993), quantile estimation (Tiwari and Wells, 1989; Nourmohammadi et al., 2014) and attribute control charts (Jafari Jozani and Mirkamali 2010, 2011).

The outline of the paper is as follows. In Section 2, we describe the RNS design and present some basic distribution theory of the random variables associated with this design when the underlying population is absolutely continuous. In Section 3, we study the construction of the exact and asymptotic RNS confidence intervals for quantiles. Several interesting theoretical results are presented in this section. We show a duality between the problem of constructing an RNS confidence interval for the p-th quantile of the underlying population with the one for the p'-th quantile of the population based on SRS, where p' is obtained as a function of p and the parameters of the RNS design (see Theorem 3.2). We compare the RNS and SRS confidence intervals based on their length and coverage probabilities and the necessary sample size to achieve the desired coverage probabilities in each design. We observe that the design parameter ζ, associated with RNS, provides a

flexible tool which enables one to construct confidence intervals with the exact desired coverage probabilities for a wide range of population quantiles without the use of randomized procedures (see Section 3.3). Section 4 provides a case study, based on the small livestock data set used in Ozturk et al. (2005) and Jafari Jozani and Johnson (2012) for both perfect and imperfect ranking situations. Finally, Section 5 provides some concluding remarks, while the Appendix is devoted to some of the proofs.

2 Preliminary Results on Distributional Properties of RNS Samples

In this section, we describe the RNS design and present some basic distribution theory of the random variables associated with it when the underlying population is absolutely continuous.

Let $K_1, K_2, \ldots$ be a sequence of independent random variables taking values in $\{1, \ldots, N\}$ with probabilities $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_N)$, where $0 \leq \rho_i \leq 1$ and $\rho_1 + \cdots + \rho_N = 1$. Furthermore, let $Z_1, Z_2, \ldots$ be a sequence or independent Bernoulli random variables with success probability $\zeta \in [0, 1]$, such that $\{K_i : i \in \mathbb{N}\}$ and $\{Z_i : i \in \mathbb{N}\}$ are independent. To obtain an RNS$(\boldsymbol{\rho}, \zeta)$ sample of size m, for each $i = 1, \ldots, m$, we perform the following steps:

1. Observe K_i and take a simple random sub-sample $X_1, \ldots, X_{K_i}$ of size K_i from the population.
2. Rank the observations in the sub-sample, either by visual inspection or some easy to measure auxiliary attribute to obtain the ordered sub-sample $X_{1:K_i}, X_{2:K_i}, \ldots, X_{K_i:K_i}$.
3. Observe Z_i and select the element $Y_i(\zeta) = (1 - Z_i)X_{1:K_1} + Z_i X_{K_i:K_i}$ for measurement.

The resulting collection $\boldsymbol{Y}_{\mathrm{RNS}(\boldsymbol{\rho}, \zeta)} = \{Y_i(\zeta) : i = 1, \ldots, m\}$ is called an RNS$(\boldsymbol{\rho}, \zeta)$ sample of size m and, throughout, $Y(\zeta)$ will denote an observation from an RNS$(\boldsymbol{\rho}, \zeta)$ design.

Remark 2.1. As an example to how the ranking in step 2 may be achieved, suppose we have an auxiliary attribute X' which is easy to measure or rank visually. Then the i-th subsample would consist of pairs $(X_1, X'_1), \ldots, (X_{K_i}, X'_{K_i})$. This subsample could then be easily ordered according to the X'_j, resulting in the "ordered" sample $(X_{[1:K_i]}, X'_{1:K_i}), \ldots, (X_{[K_i:K_i]}, X'_{K_i:K_i})$, where $X_{[j:K_i]}$ denotes the X associated with $X'_{j:K_i}$. This

may result in imperfect ranking of the X's, which is discussed further in Section 4.

We begin with a basic expectation result.

LEMMA 2.2. *Let $Y_1(\zeta), Y_2(\zeta), \ldots, Y_m(\zeta)$ be an RNS($\boldsymbol{\rho}, \zeta$) sample of size m taken from a continuous distribution with cumulative distribution function (cdf) F. Let $\boldsymbol{X}_i = (X_1, \ldots, X_{K_i})$ denote a simple random sub-sample of size K_i from F and let $\boldsymbol{U} : \mathbb{R}^m \to \mathbb{R}$ be a real valued function. For $A \subseteq \mathbb{N}$, we define*

$$V_A(\boldsymbol{X}_i, K_i) = 1_{A^c}(i) X_{1:K_i} + 1_A(i) X_{K_i:K_i},$$

where $1_A(\cdot)$ is the indicator function of A. Then, subject to the existence of the expectations involved,

$$\begin{aligned} E\left[\boldsymbol{U}(Y_1(\zeta), \ldots, Y_m(\zeta))\right] &= \sum_{A \in \mathcal{P}(S)} \zeta^{n(A)} (1-\zeta)^{m-n(A)} \\ &\quad \times E\left[\boldsymbol{U}\left(V_A(\boldsymbol{X}_1, K_1), \ldots, V_A(\boldsymbol{X}_{K_m}, K_m)\right)\right], \end{aligned}$$

where the summation extends over $\mathcal{P}(S)$, the power set of $S = \{1, 2, \ldots, m\}$, and $n(A)$ denotes the number of elements in A.

PROOF. Since $\boldsymbol{U}\big(Y_1(\zeta), \ldots, Y_m(\zeta)\big)$ may be written as

$$\sum_{A \in \mathcal{P}(S)} \Big(\prod_{j \in A} Z_j\Big)\Big(\prod_{j \in A^c} (1 - Z_j)\Big) \boldsymbol{U}\left(V_A(\boldsymbol{X}_1, K_1), \ldots, V_A(\boldsymbol{X}_{K_m}, K_m)\right),$$

independence of $\{Z_i\}$, $\{K_i\}$ and the $\{\boldsymbol{X}_i\}$ yields the desired result.

Taking $m = 1$ and U the identity map easily yields the following corollary.

COROLLARY 2.3. *Let $Y_1(\zeta), Y_2(\zeta), \ldots, Y_m(\zeta)$ be an RNS($\boldsymbol{\rho}, \zeta$) sample of size m taken from a continuous distribution function F. Subject to the existence of expectations, if $U : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[U(Y_i(\zeta))] = \zeta\, \mathbb{E}[U(X_{K_i:K_i})] + (1 - \zeta)\, \mathbb{E}[U(X_{1:K_i})].$$

In particular,

$$\mathbb{E}[Y_i(\zeta)] = \zeta \sum_{k=1}^{N} \rho_k\, \mu_{k:k} + (1 - \zeta) \sum_{k=1}^{N} \rho_k\, \mu_{1:k},$$

where $\mu_{j:k} = \mathbb{E}[X_{j:k}]$ is the expected value of the j-th order statistic in a simple random sample of size k from F.

Corollary 2.3 allows us to easily determine the distribution function of $Y_i(\zeta)$ and show that it enjoys a certain symmetry property.

COROLLARY 2.4. *Let $Y_1(\zeta), Y_2(\zeta), \ldots, Y_m(\zeta)$ be an RNS($\boldsymbol{\rho}, \zeta$) sample of size m taken from a continuous distribution F and let $\bar{F} = 1 - F$. The cdf of $Y_i(\zeta)$ is given by*

$$G(y) = \zeta \sum_{k=1}^{N} \rho_k F^k(y) + (1-\zeta) \sum_{k=1}^{N} \rho_k (1 - \bar{F}^k(y)) := \Delta_{\boldsymbol{\rho}}(F(y), \zeta), \quad \forall y \in \mathbb{R}. \tag{1}$$

Furthermore,

$$\Delta_{\boldsymbol{\rho}}(F(y), \zeta) = 1 - \Delta_{\boldsymbol{\rho}}(1 - F(y), 1 - \zeta). \tag{2}$$

PROOF. The cdf of $Y_i(\zeta)$ is easily obtained by taking $U(Y_i(\zeta)) = 1_{(-\infty, y]}(Y_i(\zeta))$, $y \in \mathbb{R}$, in Corollary 2.3. Also,

$$\begin{aligned} 1 - \Delta_{\boldsymbol{\rho}}(1 - F(y), 1 - \zeta) &= 1 - (1-\zeta) \sum_{k=1}^{N} \rho_k (1 - F(y))^k - \zeta \sum_{k=1}^{N} \rho_k (1 - F^k(y)) \\ &= \zeta \sum_{k=1}^{N} \rho_k F^k(y) + (1-\zeta) \sum_{k=1}^{N} \rho_k \left(1 - (1 - F(y))^k\right) \\ &= \Delta_{\boldsymbol{\rho}}(F(y), \zeta), \quad \forall y \in \mathbb{R}, \end{aligned}$$

which completes the proof.

In Corollary 2.4, taking $\mathbb{P}(K_i = 1) = \rho_1 = 1$ results in $G(y) = F(y)$, where G and F are the cdfs of $Y(\zeta)$ and X, respectively. Also from (1), it is easy to see that $Y_i(\zeta)$, $i = 1, \ldots, m$, are independent and identically distributed (i.i.d.) random variables from $G(\cdot)$ with a pdf $g(y) = f(y)\delta_{\boldsymbol{\rho}}(F(y), \zeta)$, where $\delta_{\boldsymbol{\rho}}(F(y), \zeta)$ is a weight function defined by

$$\delta_{\rho}(F(y), \zeta) := \zeta \sum_{k=1}^{N} k\rho_k \, F^{k-1}(y) + (1-\zeta) \sum_{k=1}^{N} k\rho_k \, \bar{F}^{k-1}(y).$$

If $\mathbb{P}(K_i \in \{1, 2\}) = 1$ for all i (i.e. $\rho_1 + \rho_2 = 1$), Corollary 2.4 yields

$$G(y) = \mathbb{P}(Y(\zeta) \le y) = F(y) + \mathbb{P}(K = 2) F(y)(1 - F(y))(2\zeta - 1),$$

which leads to the following useful result regarding the stochastic ordering of $X \sim F$ and $Y(\zeta) \sim G$.

COROLLARY 2.5. *Let F be a continues distribution and let X and $Y(\zeta)$ denote observations taken from F using SRS and RNS($\boldsymbol{\rho}, \zeta$) designs, respectively. Suppose further $\mathbb{P}(K = 2) = 1 - \mathbb{P}(K = 1) \in [0, 1]$. Then $Y(\zeta)$ is stochastically smaller, equal or larger than X according to ζ smaller, equal or larger than 0.5, respectively.*

3 RNS Confidence Intervals for Quantiles

Suppose $Y_1(\zeta), \ldots, Y_m(\zeta)$ is an RNS($\boldsymbol{\rho}, \zeta$) sample of size m from F. Our goal is to obtain a confidence interval for $Q_{X,p} = F^{-1}(p)$, the p-th quantile of F $(0 < p < 1)$. Let $Y_{1:m}(\zeta), \ldots, Y_{m:m}(\zeta)$ represent the ordered observations obtained from RNS($\boldsymbol{\rho}, \zeta$). An RNS-based confidence interval for $Q_{X,p}$, with the nominal confidence coefficient $100(1-\gamma)\%$ is an interval $[Y_{r:m}(\zeta), Y_{s:m}(\zeta)]$ such that

$$\mathbb{P}\big(Y_{r:m}(\zeta) \le Q_{X,p} \le Y_{s:m}(\zeta)\big) = 1 - \gamma. \tag{3}$$

Exact equality in (3) is seldom achievable due to the discrete nature of the probability involved and, baring the use of a randomized procedure, we will have to be content in constructing an interval that satisfies

$$\mathbb{P}\big(Y_{r:m}(\zeta) \le Q_{X,p} \le Y_{s:m}(\zeta)\big) \ge 1 - \gamma. \tag{4}$$

In some applications we are interested in the (one sided) upper confidence limit, $Y_{s:m}(\zeta)$, where

$$s = \inf\Big\{i : \mathbb{P}(Q_{X,p} \le Y_{i:m}(\zeta)) \ge 1 - \gamma\Big\},$$

or the lower confidence limit, $Y_{r:m}(\zeta)$, where

$$r = \inf\Big\{i : \mathbb{P}(Y_{i:m}(\zeta) \le Q_{X,p}) \ge 1 - \gamma\Big\}.$$

In order to construct a $100(1 - \gamma)\%$ RNS-based confidence interval for $Q_{X,p}$, let M denote the number of $Y_i(\zeta)$'s which are less than or equal to $Q_{X,p}$, so that $M \sim \text{Bin}(m, \Delta_{\boldsymbol{\rho}}(p, \zeta))$. It is easy to show that

$$\begin{aligned}\mathbb{P}\big(Y_{r:m}(\zeta) \le Q_{X,p}\big) &= \mathbb{P}(M \ge r) \\ &= \sum_{j=r}^{m} \binom{m}{j} \Delta_{\boldsymbol{\rho}}^{j}(p, \zeta)\big(1 - \Delta_{\boldsymbol{\rho}}(p, \zeta)\big)^{m-j} \\ &= I_{\Delta_{\boldsymbol{\rho}}(p,\zeta)}(r, m - r + 1),\end{aligned}$$

where $I_q(a, b)$ is the incomplete beta function defined as

$$I_q(a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \int_0^q t^{a-1}(1 - t)^{b-1} dt, \;\; 0 < q < 1, \;\; a, b > 0.$$

The coverage probability associated with $[Y_{r:m}(\zeta), Y_{s:m}(\zeta)]$ is hence given by

$$\begin{aligned}\mathbb{P}\big(Y_{r:m}(\zeta) \le Q_{X,p} \le Y_{s:m}(\zeta)\big) &= \sum_{j=r}^{s-1} \binom{m}{j} \Delta_{\boldsymbol{\rho}}^{j}(p,\zeta)\big(1-\Delta_{\boldsymbol{\rho}}(p,\zeta)\big)^{m-j} \\ &= I_{\Delta_{\boldsymbol{\rho}}(p,\zeta)}(r, m-r+1) - I_{\Delta_{\boldsymbol{\rho}}(p,\zeta)}(s, m-s+1). \end{aligned} \tag{5}$$

Note that this coverage probability depends only on ζ, $\boldsymbol{\rho}$ and p, and not on F. Hence, $[Y_{r:m}(\zeta), Y_{s:m}(\zeta)]$ is a distribution-free RNS-based confidence interval for $Q_{X,p}$.

REMARK 3.1. If $X_1, \ldots, X_m$ are i.i.d. from F, then the coverage probability of the interval $[X_{r:m}, X_{s:m}]$ for $Q_{X,p}$ is given by

$$\begin{aligned}\mathbb{P}(X_{r:m} \le Q_{X,p} \le X_{s:m}) &= \sum_{j=r}^{s-1} \binom{m}{j} p^j (1-p)^{m-j} \\ &= I_p(r, m-r+1) - I_p(s, m-s+1), \end{aligned} \tag{6}$$

which is obtained from (5) by taking $\rho_1 = \mathbb{P}(K=1) = 1$ (see David and Nagaraja, 2003).

Recall that $Y(\zeta) \sim G(\cdot) = \Delta_{\boldsymbol{\rho}}(F(\cdot), \zeta)$. The following theorem establishes a one to one correspondence between the quantiles of G and the quantiles of F, which allows results for RNS based procedures to be obtained by making use of well established results for SRS designs.

THEOREM 3.2. *Suppose F is a continuous cdf and let X and Y denote observations obtained from F using SRS and RNS($\boldsymbol{\rho}, \zeta$) designs, respectively. Then $Q_{X,p} = Q_{Y,\Delta_{\boldsymbol{\rho}}(p,\zeta)}$, where $Q_{X,p}$ is the p-th quantile of X and $Q_{Y,\Delta_{\boldsymbol{\rho}}(p,\zeta)}$ is the $\Delta_{\boldsymbol{\rho}}(p,\zeta)$-th quantile of Y.*

PROOF. Note that $\Delta_{\boldsymbol{\rho}}(p,\zeta)$ is the cdf of $Y(\zeta)$ at $Q_{X,p}$ and hence it is a non-decreasing function of p. The result follows from the definition of quantiles and the following equalities:

$$\begin{aligned} Q_{X,p} &= \inf\{x : F(x) \ge p\} \\ &= \inf\{x : \Delta_{\boldsymbol{\rho}}(F(x),\zeta) \ge \Delta_{\boldsymbol{\rho}}(p,\zeta)\} \\ &= \inf\{x : G(x) \ge \Delta_{\boldsymbol{\rho}}(p,\zeta)\} \\ &= Q_{Y,\Delta_{\boldsymbol{\rho}}(p,\zeta)}. \end{aligned}$$

Let $U_1, \ldots, U_m$ be a simple random sample of size m from a uniform$[0,1]$ distribution and let $U_{i:m}$, $i = 1, \ldots, m$ denote the corresponding order statistics. Theorem 3.2, and the fact that $G(Y_{i:m}(\zeta)) \stackrel{d}{=} U_{i:m}$, yields

$$\begin{aligned}\mathbb{P}(Y_{r:m}(\zeta) \le Q_{X,p} \le Y_{s:m}(\zeta)) \ge 1-\gamma \Leftrightarrow \mathbb{P}(Y_{r:m}(\zeta) \le Q_{Y,\Delta_\rho(p,\zeta)} \\ \le Y_{s:m}(\zeta)) \ge 1-\gamma \Leftrightarrow \mathbb{P}(U_{r:m} \le \Delta_\rho(p,\zeta) \le U_{s:m}) \ge 1-\gamma. \quad (7)\end{aligned}$$

Thus, to construct a non-parametric RNS-based confidence interval $[Y_{r:m}(\zeta), Y_{s:m}(\zeta)]$ for $Q_{X,p}$ satisfying (4), one should, when it exists, choose (r,s) such that

$$(r,s) = \underset{(u,v)\in S}{\operatorname{argmin}}\{\, \mathbb{P}(U_{u:m} \le p' \le U_{v:m}) : \mathbb{P}(U_{u:m} \le p' \le U_{v:m}) \ge 1-\gamma \,\}, \quad (8)$$

where $p' = \Delta_{\boldsymbol{\rho}}(p,\zeta)$, and $U_{1:m}, \ldots, U_{m:m}$ are the order statistics from a simple random sample of size m from a uniform$[0,1]$ distribution. Solution(s) to (8) can be easily found using a table of appropriate binomial probabilities in the small sample case. Approximate solutions to (8) are discussed in the next section. In the event that (8) has multiple solutions, one would choose a solution that minimizes $s-r$, since this would minimize the expected length of the interval, which is given by

$$\mathbb{E}[U_{s:m} - U_{r:m}] = \frac{s-r}{m+1}.$$

The following lemma shows that confidence intervals for $Q_{X,p}$ based on an RNS$(\boldsymbol{\rho}, \zeta)$ design may be reflected to obtain results for $Q_{X,1-p}$ based on an RNS$(\boldsymbol{\rho}, 1-\zeta)$ design.

LEMMA 3.3. *Suppose $Q_{X,p}$ is the p-th quantile of the continuous cdf F, where $p \in [0,1]$. Let $Y_{1:m}(\zeta) \le \ldots \le Y_{m:m}(\zeta)$ be the order statistics associated with a sample of size m, obtained from F, using the RNS$(\boldsymbol{\rho}, \zeta)$ design. Then,*

(a) $[Y_{r:m}(\zeta), Y_{s:m}(\zeta)]$ is an RNS-based confidence interval of level $100(1-\gamma)\%$ for $Q_{X,p}$ if and only if $[Y_{m-s+1:m}(1-\zeta), Y_{m-r+1:m}(1-\zeta)]$ is the RNS confidence interval for $Q_{X,1-p}$, i.e.,

$$\begin{aligned}&\mathbb{P}\left(Y_{r:m}(\zeta) \le Q_{X,p} \le Y_{s:m}(\zeta)\right) \ge 1-\gamma \\ &\Leftrightarrow \mathbb{P}\left(Y_{m-s+1:m}(1-\zeta) \le Q_{X,1-p} \le Y_{m-r+1:m}(1-\zeta)\right) \ge 1-\gamma.\end{aligned}$$

(b) $Y_{s:m}(\zeta)$ *is the upper RNS-based confidence limit of level* $100(1-\gamma)\%$ *for* $Q_{X,p}$ *if and only if* $Y_{m-s+1:m}(1-\zeta)$ *is the lower RNS-based confidence limit of level* $1-\gamma$ *for* $Q_{X,1-p}$*, i.e.,*

$$\mathbb{P}\left(Q_{X,p} \le Y_{s:m}(\zeta)\right) \ge 1-\gamma \Leftrightarrow \mathbb{P}\left(Y_{m-s+1:m}(1-\zeta) \le Q_{X,1-p}\right) \ge 1-\gamma.$$

Proof. Using (2) and the coverage probability obtained in (5), we readily have

$$\begin{aligned}
&\mathbb{P}\left(Y_{r:m}(\zeta) \le Q_{X,p} \le Y_{s:m}(\zeta)\right) \ge 1-\gamma \\
&\quad\Leftrightarrow \sum_{j=r}^{s-1} \binom{m}{j} \Delta_{\boldsymbol{\rho}}^{j}(p,\zeta)\left(1-\Delta_{\boldsymbol{\rho}}(p,\zeta)\right)^{m-j} \ge 1-\gamma \\
&\quad\Leftrightarrow \sum_{j=r}^{s-1} \binom{m}{j} \Delta_{\boldsymbol{\rho}}^{m-j}(1-p,1-\zeta)\left(1-\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)\right)^{j} \ge 1-\gamma \\
&\quad\Leftrightarrow \sum_{j=m-s+1}^{m-r} \binom{m}{j} \Delta_{\boldsymbol{\rho}}^{j}(1-p,1-\zeta)\left(1-\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)\right)^{m-j} \ge 1-\gamma \\
&\quad\Leftrightarrow \mathbb{P}(Y_{m-s+1:m}(1-\zeta) \le Q_{X,1-p} \le Y_{m-r+1:m}(1-\zeta)) \ge 1-\gamma.
\end{aligned}$$

Similar argument can be used to establish the result in (b).

3.1. Large Sample Theory. When the sample size is large, one can use the normal approximation to the binomial distribution to approximate r and s. For example, for a RNS-based confidence interval of level $100(1-\gamma)\%$ for $Q_{X,p}$, we could consider

$$\begin{aligned}
r &\approx m\Delta_{\boldsymbol{\rho}}(p,\zeta) - z_{\gamma/2}\sqrt{m\Delta_{\boldsymbol{\rho}}(p,\zeta)\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)}, \\
s &\approx m\Delta_{\boldsymbol{\rho}}(p,\zeta) + z_{\gamma/2}\sqrt{m\Delta_{\boldsymbol{\rho}}(p,\zeta)\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)};
\end{aligned} \tag{9}$$

where $\mathbb{P}(Z \le z_{\gamma/2}) = \gamma/2$ with $Z \sim N(0,1)$. Also, the approximate coverage probability associated with $[Y_{r:m}(\zeta), Y_{s:m}(\zeta)]$ is given by

$$\Phi\left(\frac{s - m\Delta_{\boldsymbol{\rho}}(p,\zeta)}{\sqrt{m\,\Delta_{\boldsymbol{\rho}}(p,\zeta)\,\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)}}\right) - \Phi\left(\frac{r - m\Delta_{\boldsymbol{\rho}}(p,\zeta)}{\sqrt{m\,\Delta_{\boldsymbol{\rho}}(p,\zeta)\,\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)}}\right),$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. In SRS, the normal approximation does not perform well, especially when p is far from 0.5. However, in RNS designs, we expect the normal approximation to work better even for small (large) values of p since we can normally find a ζ such that $|\Delta_{\boldsymbol{\rho}}(p,\zeta) - 1/2| < |p - 1/2|$ to shrink $\Delta_{\boldsymbol{\rho}}(p,\zeta)$ toward $1/2$, which results in the best normal approximation to the binomial distribution.

One can also obtain an RNS confidence interval for $Q_{X,p}$ using the RNS sample quantile and its asymptotic sampling distribution. To this end, let

$$G_m(y) = \frac{1}{m}\sum_{i=1}^{m} I(Y_i(\zeta) \le y),$$

be the empirical distribution function associated with an RNS sample of size m and define

$$\hat{Q}_{X,m,p} = \inf\{y : G_m(y) \ge \Delta_{\boldsymbol{\rho}}(p,\zeta)\},$$

where $p, \zeta \in [0,1]$. Suppose F is differentiable at $Q_{X,p}$ with $F'(Q_{X,p}) > 0$. For large m, it can be shown that $\sqrt{m}(\hat{Q}_{X,m,p} - Q_{X,p})$ is asymptotically normal $\mathrm{N}(0, \sigma^2_{\boldsymbol{\rho}}(p,\zeta)/\{F'(Q_{X,p})\}^2)$, where $\sigma^2_{\boldsymbol{\rho}}(p,\zeta) = \Delta_{\boldsymbol{\rho}}(p,\zeta)\Delta_{\boldsymbol{\rho}}(1-p,1-\zeta)/\delta^2_{\boldsymbol{\rho}}(p,\zeta)$. An asymptotic RNS-based confidence interval of level $100(1-\gamma)\%$ for $Q_{X,p}$ is now given by

$$\left(\hat{Q}_{X,m,p} - z_{\gamma/2}\frac{1}{\sqrt{m}}\frac{\sigma_{\boldsymbol{\rho}}(p,\zeta)}{F'(Q_{X,p})}, \hat{Q}_{X,m,p} + z_{\gamma/2}\frac{1}{\sqrt{m}}\frac{\sigma_{\boldsymbol{\rho}}(p,\zeta)}{F'(Q_{X,p})}\right). \tag{10}$$

REMARK 3.4. Equation (10) can be used to determine the m required for a specified margin of error for the approximate large sample confidence interval, provided one has an approximation for $F'(Q_{X,p})$.

3.2. Choosing ζ for Symmetric Confidence Intervals. By a symmetric confidence interval, we mean an interval of the form $[Y_{r:m}(\zeta), Y_{m-r+1:m}(\zeta)]$. Using (5) and (6), the coverage probabilities for $Q_{X,p}$ associated with symmetric RNS($\boldsymbol{\rho},\zeta$)- and SRS-based confidence intervals are both of the form of

$$\sum_{j=r}^{m-r}\binom{m}{j}\nu^j(1-\nu)^{m-j}, \quad \nu \in [0,1], \tag{11}$$

where $\nu = p$ for SRS-based confidence intervals and $\nu = \Delta_{\boldsymbol{\rho}}(p,\zeta)$ for RNS($\boldsymbol{\rho},\zeta$)-based confidence intervals.

The following lemma summarizes the behavior of (11). The proof of the first two parts are straightforward and omitted. The proof of the last part is presented in Appendix.

LEMMA 3.5. *Let* $H_{r,m}(\nu)=\sum_{j=r}^{m-r}\binom{m}{j}\nu^j(1-\nu)^{m-j}$, *where* $r\in\{1,\dots,\lfloor m/2\rfloor\}$ *and* $\nu\in[0,1]$. *Then,*

(a) $H_{r,m}(\nu)$ *is symmetric about* $\nu=1/2$, *that is,* $H_{r,m}(\nu)=H_{r,m}(1-\nu)$.

(b) $H_{r,m}(\nu)$ *is non-increasing in* r, *that is,* $H_{r+1,m}(\nu)\le H_{r,m}(\nu)$.

(c) $H_{r,m}(\nu)$ *is increasing for* $\nu<\frac{1}{2}$ *and decreasing for* $\nu>\frac{1}{2}$.

Lemma 3.5 shows that the coverage probability of an RNS-based confidence interval is greatest when $\Delta_{\boldsymbol{\rho}}(p,\zeta)=1/2$. For symmetric populations this can easily be explained using Lemma 3.7 (see below). We first need the following definition.

Definition 3.6. *We say* X *is more peaked about a known value* ν *than* $Y(\zeta)$ *if, for all* $\epsilon>0$,

$$\mathbb{P}\left(|X-\nu|>\epsilon\right)\le\mathbb{P}\left(|Y(\zeta)-\nu|>\epsilon\right).$$

The following lemma shows that an observation X from a symmetric continuous distribution F is more peaked about the median of F than an observation $Y(\zeta)$ from an RNS$(\boldsymbol{\rho},\zeta)$ design. The proof is deferred to the Appendix.

LEMMA 3.7. *Let* X *and* $Y(\zeta)$ *denote observations taken from a symmetric continuous distribution* F *using SRS and RNS*$(\boldsymbol{\rho},\zeta)$, *respectively. Then* X *is more peaked about* $Q_{X,1/2}$ *than* $Y(\zeta)$.

In light of (8), this implies that an RNS$(\boldsymbol{\rho},\zeta)$-based interval $[Y_{r:m}(\zeta), Y_{m-r+1:m}]$ will have higher coverage probability than an SRS-based interval $[X_{r:m},X_{m-r+1:m}]$ provided $|\Delta_{\boldsymbol{\rho}}(p,\zeta)-1/2|<|p-1/2|$. The following lemma and theorem establishes that this is always possible, provided $p\neq1/2$. The proof of the following is deferred to the Appendix.

LEMMA 3.8. *Let*

$$\alpha_{\boldsymbol{\rho}}(p)=\sum_{k=1}^{N}\rho_k p^k\quad\textit{and}\quad\beta_{\boldsymbol{\rho}}(p)=1-\alpha_{\boldsymbol{\rho}}(1-p),\tag{12}$$

where $\rho_i,p\in(0,1)$, $\rho_1+\cdots+\rho_N=1$ *and, by construction,* $\alpha_{\boldsymbol{\rho}}(p)\le p\le\beta_{\boldsymbol{\rho}}(p)$. *Define*

$$\psi_{\boldsymbol{\rho}}(p)=\frac{\beta_{\boldsymbol{\rho}}(p)-\frac{1}{2}}{\beta_{\boldsymbol{\rho}}(p)-\alpha_{\boldsymbol{\rho}}(p)}\quad\textit{and}\quad\phi_{\boldsymbol{\rho}}(p)=\frac{\beta_{\boldsymbol{\rho}}(p)-p}{\beta_{\boldsymbol{\rho}}(p)-\alpha_{\boldsymbol{\rho}}(p)}.\tag{13}$$

Then

(a) $\psi_{\boldsymbol{\rho}}(p)$ *is non-decreasing in* p, *with* $p \in (0,1)$. *Also,* $\psi_{\boldsymbol{\rho}}(p) < 0$ *if and only if* $\beta_{\boldsymbol{\rho}}(p) < \frac{1}{2}$ *and* $\psi_{\boldsymbol{\rho}}(p) > 1$ *if and only if* $\alpha_{\boldsymbol{\rho}}(p) > \frac{1}{2}$.

(b) $\phi_{\boldsymbol{\rho}}(p)$ *is non-increasing in* p *and* $\phi_{\boldsymbol{\rho}}(p) \in (0,1)$ *for all* $p \in (0,1)$.

(c) $\psi_{\boldsymbol{\rho}}(\frac{1}{2}) = \phi_{\boldsymbol{\rho}}(\frac{1}{2}) = \frac{1}{2}$ *and, when* $\rho_1 = 1 - \rho_2$, $\phi_{\boldsymbol{\rho}}(p) = 1/2$ *for all* $p \in (0,1)$.

For a fixed and known $\boldsymbol{\rho}$, we can now characterize the range of ζ that results in the coverage probability of RNS-based intervals to be greater that the coverage probability of their SRS counterparts.

THEOREM 3.9. *Suppose* $[X_{r:m}, X_{m-r+1:m}]$ *is a* $100(1-\gamma)\%$ *SRS-based confidence interval for the* p*-th quantile* $Q_{X,p}$ *of a (continuous) cdf* F *and let* $[Y_{r:m}, Y_{m-r+1:m}]$ *be an RNS*$(\boldsymbol{\rho}, \zeta)$*-based confidence interval for* $Q_{X,p}$.

(a) *For any* $p \in (0, 1/2)$, *the coverage probability associated with* $[Y_{r:m}, Y_{m-r+1:m}]$, *is greater than* $1-\gamma$ *if* $\max\{0, \psi_{\boldsymbol{\rho}}(p)\} \le \zeta < \phi_{\boldsymbol{\rho}}(p)$, *and it is maximized* $\zeta^* = \max\{0, \psi_{\boldsymbol{\rho}}(p)\}$.

(b) *For any* $p \in (1/2, 1)$, *the coverage probability associated with* $[Y_{r:m}, Y_{m-r+1:m}]$ *is greater than* $1-\gamma$ *if* $\phi_{\boldsymbol{\rho}}(p) < \zeta \le \min\{1, \psi_{\boldsymbol{\rho}}(p)\}$, *and it is maximized at* $\zeta^* = \min\{1, \psi_{\boldsymbol{\rho}}(p)\}$.

(c) *For* $p = 1/2$ *the coverage probability associated with RNS-based confidence interval* $[Y_{r:m}, Y_{m-r+1:m}]$ *reaches its maximum value when* $\zeta = 1/2$, *and the maximum coverage probability is equal to the coverage probability associated with its SRS counterpart.*

PROOF. Using (5) and Lemma 3.5, $\mathbb{P}(Y_{r:m} \le Q_{X,p} \le Y_{m-r+1:m}) = H_{r,m}(\Delta_{\boldsymbol{\rho}}(p,\zeta))$. The coverage probability of RNS confidence interval is bigger than its SRS counterpart when $p < \Delta_{\boldsymbol{\rho}}(p,\zeta) \le 1/2$ for $p < 1/2$ and $1/2 \le \Delta_{\boldsymbol{\rho}}(p,\zeta) < p$ for $p > 1/2$. Now, the results follow upon straightforward calculations using Lemma 3.8 and the expression given for $\Delta_{\boldsymbol{\rho}}(p,\zeta)$ in (1).

Now, using Theorem 3.9 we can present the following algorithm for constructing an efficient symmetric confidence interval for $Q_{X,p}$ with RNS$(\boldsymbol{\rho}, \zeta)$ design. When $p \neq 1/2$,

1. Calculate ζ^* using Theorem 3.9 and $p^* = \Delta_{\boldsymbol{\rho}}(p, \zeta^*)$ using (1).
2. Let $r = \max\{u : H_{u,m}(p^*) \ge 1-\gamma\}$.

The above algorithm generates the RNS-based confidence interval $[Y_{r:m}(\zeta^*), Y_{m-r+1}(\zeta^*)]$ which has higher coverage probability than $[X_{r:m}, X_{m-r+1:m}]$ under SRS design (provided $p \neq 1/2$). Figure 1 shows the range of the values of ζ which results in an improved RNS(ρ, ζ) over SRS for estimating $Q_{X,p}$, $p \in (0,1)$.

For a given RNS$(\boldsymbol{\rho}, \zeta)$ design, the range of p's for which RNS$(\boldsymbol{\rho}, \zeta)$ improves, in terms of coverage probability, over SRS for estimating $Q_{X,p}$ is given in the following lemma, the proof of which is left to the reader.

LEMMA 3.10. *Consider an RNS($\boldsymbol{\rho}$, ζ) design with $\zeta \in [0,1]$. Let p_1 be the solution to $\psi_{\boldsymbol{\rho}}(p_1) = \zeta$, let p_2 be the solution to $\phi_{\boldsymbol{\rho}}(p_2) = \zeta$, and, by convention, let $[a,b] = \emptyset$ whenever $b < a$.*

(a) If $\zeta < 1/2$, then RNS$(\boldsymbol{\rho}, \zeta)$ improves, in terms of coverage probability, over SRS for estimating $Q_{X,p}$ for all then for all $p \in [0, p_1] \cup [p_2, 1]$.

(b) If $\zeta > 1/2$, then RNS$(\boldsymbol{\rho}, \zeta)$ improves, in terms of coverage probability, over SRS for estimating $Q_{X,p}$ for all then for all $p \in [0, p_2] \cup [p_1, 1]$.

(c) If $\zeta = 1/2$, then RNS$(\boldsymbol{\rho}, \zeta)$ improves, in terms of coverage probability, over SRS for estimating $Q_{X,p}$ for all $p \in [0,1] \setminus \{\frac{1}{2}\}$.

The optimum ζ^* given in Theorem 3.9 is chosen to minimize $|\Delta_{\boldsymbol{\rho}}(p, \zeta) - 1/2|$ (that is, to shrink $\Delta_{\boldsymbol{\rho}}(p, \zeta)$ toward $1/2$). The following lemma tells us precisely when $\Delta_{\boldsymbol{\rho}}(p, \zeta^*) = 1/2$, and hence, when the performance of RNS$(\boldsymbol{\rho}, \zeta^*)$ for estimating $Q_{X,p}$ is equivalent to the performance of SRS for estimating $Q_{X,1/2}$.

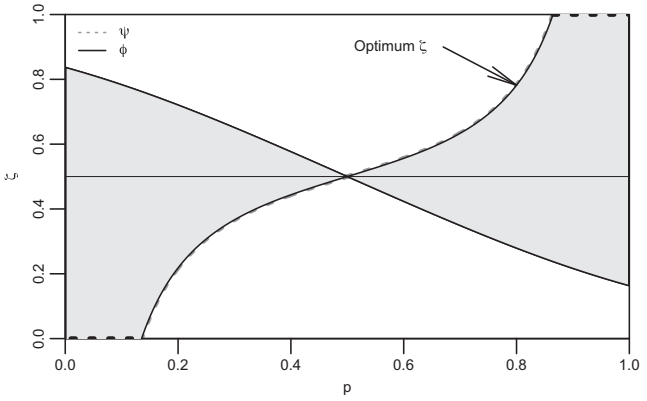

Figure 1: The area of ζ which improves RNS(ρ, ζ) over SRS for $p \in (0,1)$

LEMMA 3.11. *Let $\boldsymbol{\rho}$ be fixed (and known) and consider an RNS($\boldsymbol{\rho}, \zeta^*$) design, where ζ^* is the optimum value of ζ from Theorem 3.9. Then $\Delta_{\boldsymbol{\rho}}(p, \zeta^*) = 1/2$ for all $p \in (1-p', p')$ where p' is the (unique) solution of $\alpha_{\boldsymbol{\rho}}(p) = 1/2$.*

REMARK 3.12. The above results and algorithm assume a symmetric confidence interval of the form $[Y_{r:m}(\zeta), Y_{m-r+1:m}(\zeta)]$. One can also construct non-symmetric intervals using various methods. For example, use the asymptotic procedures in Section 3.1, or one could choose (r, s) such that

$$\begin{aligned} r &= \max\{u : \Pr(Y_{u:m}(\zeta) < Q_{X,p}) < \gamma/2\} \quad \text{and} \\ s &= \min\{u : \Pr(Y_{u:m}(\zeta) > Q_{X,p}) < \gamma/2\}, \end{aligned}$$

which would guarantee a coverage probability of at least $1-\gamma$, with approximately probability $\gamma/2$ in each tail. In the sequel, we will refer to this as the *equal-tail* interval.

3.3. On the Optimal Choice of ζ. In the previous section, we considered the range of ζ for which, given r, the coverage probability of the symmetric RNS($\boldsymbol{\rho}, \zeta$)-based interval $[Y_{r:m}(\zeta), Y_{m-r+1:m}(\zeta)]$ is larger that the corresponding SRS based interval $[X_{r:m}, X_{m-r+1:m}]$. We now consider how to choose $(r^{\mathrm{o}}, \zeta^{\mathrm{o}})$ so that

$$\Pr\big(Q_{X,p} \in [Y_{r^{\mathrm{o}}:m}(\zeta^{\mathrm{o}}), Y_{m-r^{\mathrm{o}}+1:m}(\zeta^{\mathrm{o}})]\big) = 1-\gamma.$$

In the SRS case, the difficulty lies in the discreteness of the binomial distribution and the fact that, for any given p, a solution in r to the equation $H_{r,m}(p) = 1-\gamma$ rarely exists. In the RNS($\boldsymbol{\rho}, \zeta$) design however, the additional (continuous) design parameter ζ may allow many solutions (in r and ζ) to the equation

$$H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, \zeta)) = 1-\gamma.$$

In this section, we will describe a method to obtain the best choice of (r, ζ) for estimating $Q_{X,p}$ for some fixed p.

Recall that, by Theorem 3.9, ζ^* maximizes $H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, \zeta))$ (independently of r). Therefore, for fixed p and r, provided that

$$\begin{aligned} H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, \zeta^*)) &\geq 1-\gamma \quad \text{and either} \quad H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, 0)) < 1-\gamma \quad \text{or} \\ H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, 1)) &< 1-\gamma, \end{aligned}$$

there exists at least one solution, say ζ_r^{o}, such that $H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, \zeta_r^{\mathrm{o}})) = 1-\gamma$. To this end, define $\mathfrak{S}_p$ to be the set of all solutions in r and ζ for a given p:

$$\mathfrak{S}_p = \{\, (r, \zeta) : H_{r,m}(\Delta_{\boldsymbol{\rho}}(p, \zeta)) = 1-\gamma \,\}.$$

If $\mathfrak{S}_p$ is not empty, we select a solution $(r^{\mathrm{o}}, \zeta^{\mathrm{o}}) \in \mathfrak{S}_p$ corresponding to a solution with the largest r. If more than one solution exists for the maximum r, the choice is arbitrary and can be made based on other criteria. For the purposes of this paper, when more than one solution exists for the largest r, we use the one with the larger ζ. That is, we choose

$$r^{\mathrm{o}} = \max\{r : (r, \cdot) \in \mathfrak{S}_p\} \qquad \text{and} \qquad \zeta^{\mathrm{o}} = \max\{\zeta : (r^{\mathrm{o}}, \zeta) \in \mathfrak{S}_p\}.$$

When a solution does not exist, we suggest taking $\zeta_o = \zeta^*$ and, when it exists,

$$r^{\mathrm{o}} = \max\{r : H_{r,m}(p) \geq 1 - \gamma\},$$

which guarantees a coverage probability of at least $1 - \gamma$.

A sufficient condition for $\mathfrak{S}_p$ not to be empty is

$$\begin{aligned} H_{1,m}(\Delta_{\boldsymbol{\rho}}(p, \zeta^*)) &\geq 1 - \gamma \quad \text{and either} \quad H_{1,m}(\Delta_{\boldsymbol{\rho}}(p, 0)) < 1 - \gamma \quad \text{or} \\ H_{1,m}(\Delta_{\boldsymbol{\rho}}(p, 1)) &< 1 - \gamma, \end{aligned}$$

which, for fixed $\boldsymbol{\rho}$, may be used to determine the minimum m required for such a solution to exist.

3.4. An Illustrative Example. To highlight the differences between RNS- and SRS-based symmetric confidence intervals we consider four distributions on K. In each case, the K_j are i.i.d. random variables with support $\{1, \ldots, 4\}$. The distributions considered are

$$\begin{aligned} &\boldsymbol{\rho}_0 = (1, 0, 0, 0), \boldsymbol{\rho}_1 = (0.818, 0.164, 0.017, 0.001), \\ &\boldsymbol{\rho}_2 = (0.1, 0.2, 0.3, 0.4), \text{ and } \boldsymbol{\rho}_3 = (0, 0, 0, 1). \end{aligned}$$

The distribution $\boldsymbol{\rho}_0$ corresponds to SRS (see the discussion after Corollary 2.4) and, for this design, we have $\Delta_{\boldsymbol{\rho}_0}(p, \zeta) = p$. The distribution $\boldsymbol{\rho}_1$ corresponds to the case where $K_i - 1$ are i.i.d. truncated Poisson($\lambda = 0.2$) random variables with support $\{0, 1, 2, 3\}$, as suggested by Jafari Jozani and Johnson (2012) for estimating the population total associated with the variable of interest in the case study considered in Section 4. By choosing $\boldsymbol{\rho}_2$ and $\boldsymbol{\rho}_3$ we allow RNS designs which favor larger subsample sizes more than $\boldsymbol{\rho}_0$ and $\boldsymbol{\rho}_1$. The choice of $\boldsymbol{\rho}_3$ corresponds to the RNS design with a fixed set size $K_j = 4$, $i = 1, \ldots, 4$. Throughout these examples, we use a sample size of $m = 45$, which is used in the case study to follow.

Figure 2 (left panel) plots $\Delta_{\boldsymbol{\rho}_i}(p, \zeta_i^*)$ where, for each $i = 0, \ldots, 3$ and $p \in (0, 1)$, ζ_i^* was chosen according to Theorem 3.9. It illustrates that $|\Delta_{\boldsymbol{\rho}_i}(p, \zeta_i^*) - 1/2| < |p - 1/2|$ for $i = 1, 2, 3$, and $\Delta_{\boldsymbol{\rho}_0}(p, \zeta^*) = p$ corresponds to

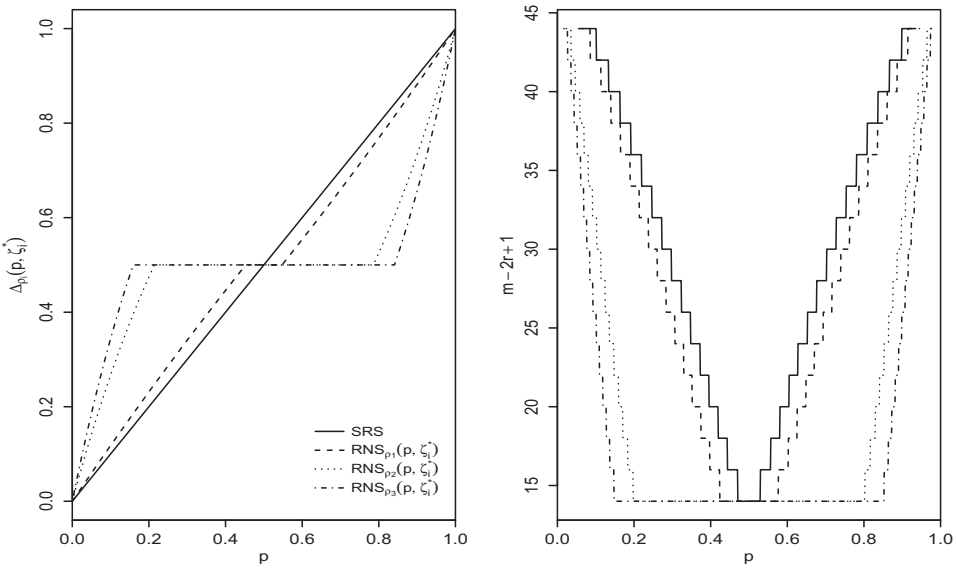

Figure 2: Comparison of p and $\Delta_{\boldsymbol{\rho}}(p, \zeta^*)$ and comparison of $m - 2r + 1$ in SRS, RNS$(\boldsymbol{\rho}_1, \zeta_1^*)$, RNS$(\boldsymbol{\rho}_2, \zeta_2^*)$, and RNS$(\boldsymbol{\rho}_3, \zeta_3^*)$, where ζ_i^* is the optimum value of ζ as functions of $\boldsymbol{\rho}_i$, $i = 1, 2, 3$, $m = 45$ and r is chosen as the largest value for which the coverage probability exceeds the nominal 0.95 level

the SRS case, illustrating the result in Lemma 3.8. It also illustrates that the amount of shrinkage toward 1/2 depends on $\boldsymbol{\rho}$, where distributions favoring larger subsample sizes tend to shrink toward 1/2 more quickly. Thus, except for inferences about the median, symmetric confidence intervals for p based on RNS$(\boldsymbol{\rho}, \zeta^*)$ designs with $\rho_1 < 1$ and the optimum ζ^* should produce shorter confidence intervals than the corresponding SRS design. Notice also that the region for which $\Delta_{\boldsymbol{\rho}_i}(p, \zeta_i^*) = 1/2$ corresponds to the region given in Lemma 3.11. For example $\alpha_{\boldsymbol{\rho}_1}(0.8408964) = 1/2$ and hence $\Delta_{\boldsymbol{\rho}_1}(p, \zeta_1^*) = 1/2$ for all $p \in (0.1591036, 0.8408964)$.

For symmetric 95 % confidence intervals of the form $[Y_{r:m}(\zeta), Y_{m-r+1}(\zeta)]$, Fig. 2 (right panel) plots the value of $(m - r + 1) - r = m - 2r + 1$ against p for designs RNS$(\boldsymbol{\rho}_i, \zeta_i^*)$, and shows that for the RNS designs with $\rho_1 < 1$ $(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \boldsymbol{\rho}_3)$, $m - 2r + 1$ is less than in SRS.

Figure 3 shows the actual coverage probability for symmetric and equal-tail 95 % confidence intervals for $Q_{X,p}$ for $p \in (0, 1)$ using the RNS$(\boldsymbol{\rho}_i, \zeta_i^*)$ designs as described in the comments after Theorem 3.9. One can easily see that the coverage probabilities associated with RNS confidence intervals are closer to the nominal 95 % confidence level than those based on SRS design.

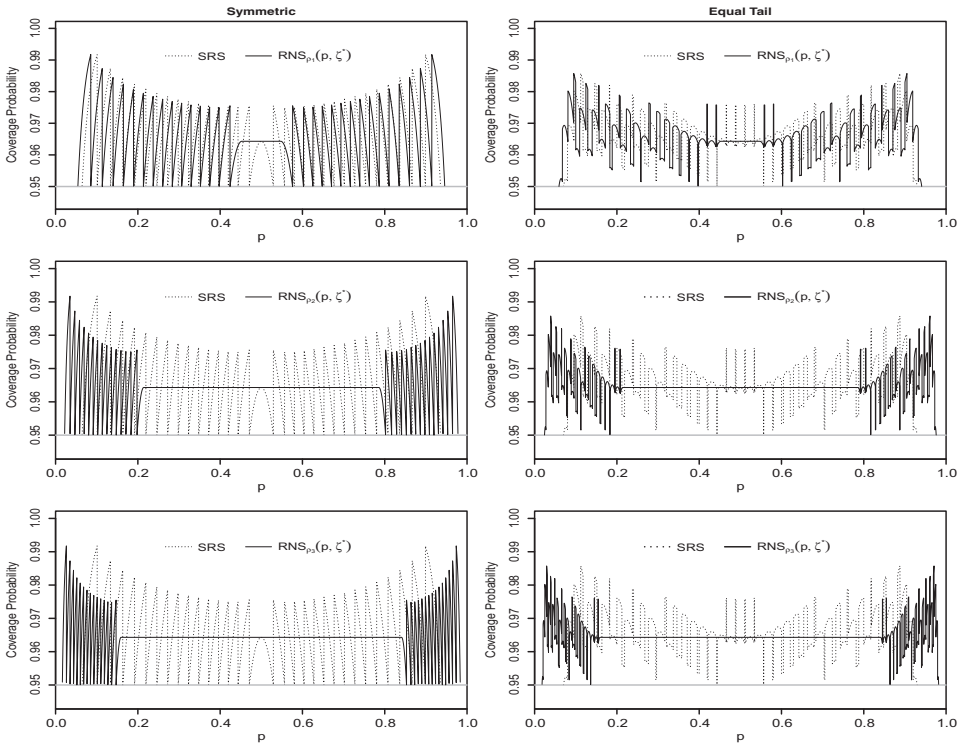

Figure 3: Comparison of the coverage probabilities of exact confidence intervals for $p \in [0,1]$ in SRS, RNS$(\boldsymbol{\rho}_1, \zeta_1^*)$, RNS$(\boldsymbol{\rho}_2, \zeta_2^*)$, and RNS$(\boldsymbol{\rho}_3, \zeta_3^*)$, where ζ_i^* is the optimum value of ζ as functions of $\boldsymbol{\rho}_i$, $i = 1, 2, 3$. The nominal confidence level is 0.95

The performance of RNS$(\boldsymbol{\rho}_i, \zeta_i^*)$ becomes better with i, $i = 1, 2, 3$. Recall that r (and s) are chosen so that the coverage probabilities are greater than or equal to $1 - \gamma$. Since this nominal level is rarely achieved in practice, the coverage probabilities in Fig. 3 are slightly higher than the nominal level $1 - \gamma$.

Figure 4 compares the coverage probabilities of the interval $[Y_{1:m}(\zeta), Y_{m:m}(\zeta)]$ for RNS$(\boldsymbol{\rho}_i, \zeta_i^*)$ designs $i = 1, 2, 3$, and the SRS design and various sample sizes m. Again, we see that RNS designs with more weight placed on larger subsample sizes perform much better for the small quantiles ($p's$) examined here. The actual sample sizes required for the interval $[Y_{1:m}(\zeta), Y_{m:m}(\zeta)]$ to achieve at least a 95 % level of confidence of containing $Q_{X,p}$ are given in Table 1. It is interesting to note that the necessary sample size for $[Y_{1:m}(\zeta_3^*), Y_{m:m}(\zeta_3^*)]$ under RNS$(\boldsymbol{\rho}_3, \zeta_3^*)$ design to achieve the desired

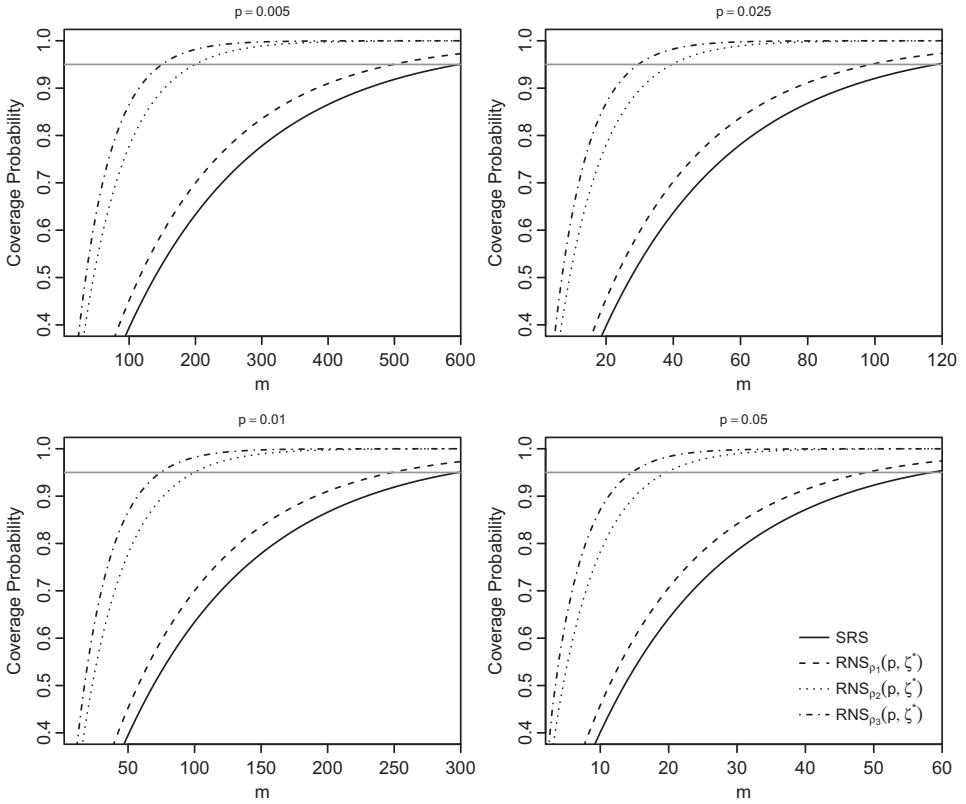

Figure 4: Comparison of the coverage probabilities of confidence intervals for $p = 0.005, 0.01, 0.025$ and 0.05 in SRS, RNS$(\boldsymbol{\rho}_1, \zeta_1^*)$, RNS$(\boldsymbol{\rho}_2, \zeta_2^*)$, and RNS$(\boldsymbol{\rho}_3, \zeta_3^*)$, ζ_i^* is the optimum value of ζ calculated as functions of $\boldsymbol{\rho}_i$, $i = 1, 2, 3$. The nominal confidence level is 0.95

95 % confidence level, is about $1/\max\{K_j\}$ times the one needed under SRS design. This is a very important observation and shows that RNS design could be very cost-effective sampling method compared with the commonly used SRS design for interval estimation of upper or lower quantiles of the underlying population.

Figure 5 compares the average coverage probabilities for the interval $[Y_{1:m}(\zeta_i^*), Y_{m:m}(\zeta_i^*)]$ for the RNS$(\boldsymbol{\rho}_i, \zeta_i^*)$ $(i = 1, 2, 3)$ and the SRS design, the average being taken over $p = 0.01(0.001)0.9$ and ζ_i^* is the optimum value of ζ as a function of $\boldsymbol{\rho}_i$ and p. The RNS designs which favor larger subsample sizes $(i = 2, 3)$ reach the nominal 95 % level much quicker than the SRS and RNS$(\boldsymbol{\rho}_1, \zeta_1^*)$ designs.

Table 1: Minimum values of sample size m in 95 % SRS and RNS$(\boldsymbol{\rho}_i, \zeta_i^*)$ confidence intervals for $Q_{X,p}$, $i = 1, 2, 3$, when $p = 0.005, 0.01, 0.025$ and 0.05 with known $\boldsymbol{\rho}_i$ and optimum values of ζ calculated as functions of $\boldsymbol{\rho}_i$

p	SRS	RNS$(\boldsymbol{\rho}_1, \zeta_1^*)$	RNS$(\boldsymbol{\rho}_2, \zeta_2^*)$	RNS$(\boldsymbol{\rho}_3, \zeta_3^*)$
0.005	598	499	200	150
0.010	299	249	100	75
0.025	119	99	40	30
0.050	59	49	20	15

4 A Case Study

In this section, we use a data set containing the birth weight and seven-month weight of 224 lambs along with the mother's weight at time of mating, collected at the Research Farm of Ataturk University, Erzurum, Turkey. Ozturk et al. (2005) as well as Jafari Jozani and Johnson (2012) used this data set to study the performance of ranked set sampling in estimating the mean and the total values of the seven-month weight of these lambs. They state that the measurement of young sheep can be labor intensive due to their active nature, and measurement errors can be inflated due to this activity. Here, we treat these 224 records as our population, with the goal

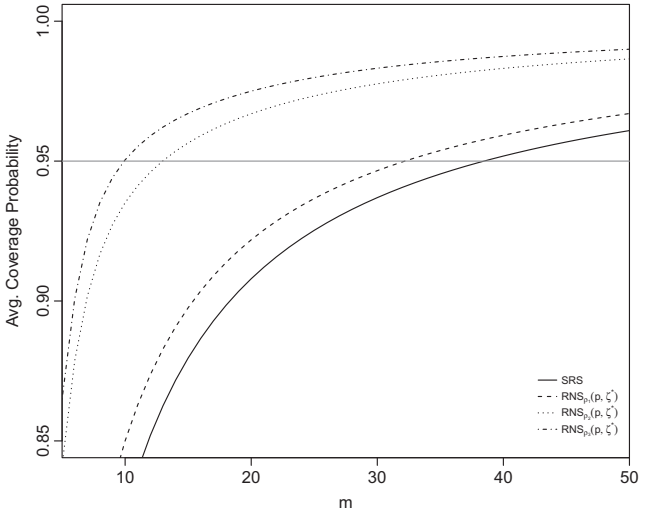

Figure 5: Comparison of the average coverage probability of the interval $[Y_{1:m}(\zeta_i^*), Y_{m:m}(\zeta_i^*)]$ in SRS, RNS$(\boldsymbol{\rho}_1, \zeta_1^*)$, RNS$(\boldsymbol{\rho}_2, \zeta_2^*)$, and RNS$(\boldsymbol{\rho}_3, \zeta_3^*)$, where ζ_i^* is the optimum value of ζ calculated as functions of p and $\boldsymbol{\rho}_i$, $i = 1, 2, 3$

of estimating the quantiles of the weight distribution of these 224 lambs at seven-month. We consider both perfect and imperfect ranking situations. For the perfect ranking scenario, ranking is done based on the weight of lambs at seven-month. For the imperfect ranking, we consider two cases. In the first case (Imperfect 1), using the mother's weight at time of mating and the birth weight of the lambs we fit a multiple linear regression model between these variables and the weight of lambs at seven-month and then we use this model to obtain the predicted values of the weight of the lambs at seven-month in each set using the available and easily obtain concomitant variables. The Kendall's τ between the seven-month weight values and their predicted values based on the fitted model is about 0.7 which is a moderate value. In the second case (Imperfect 2), we perform the ranking process based on the mother's weight at time of mating which results in an small Kendal's τ of 0.41 between the lambs weight at seven-month and mother's weight at the time of mating.

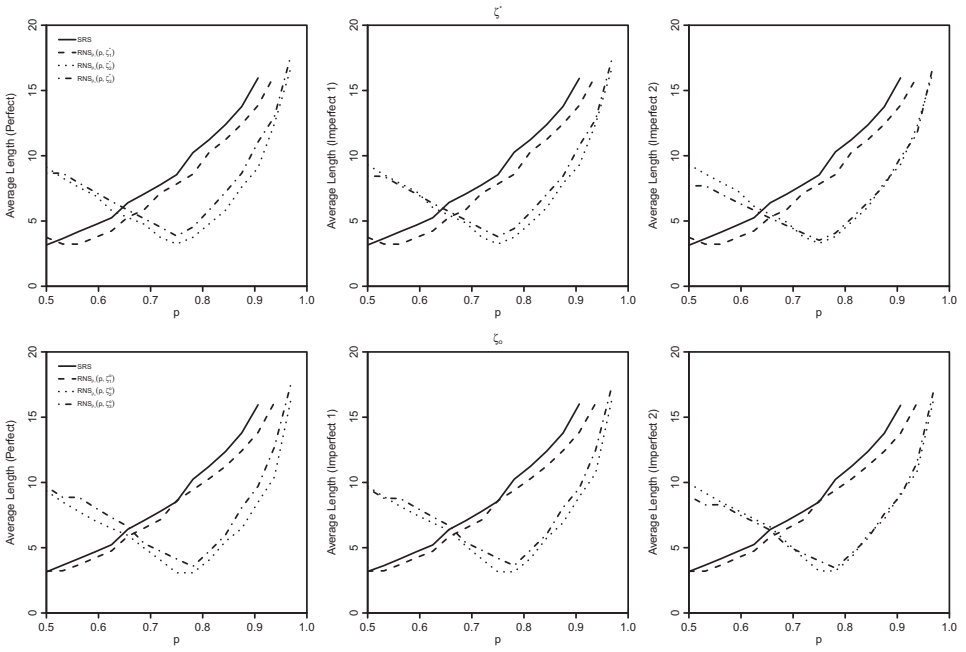

Figure 6: Average interval length for 5,000 confidence intervals using $\zeta^*_{i,75}$ (*top row*) and $\zeta^{\circ}_{i,0.75}$ (*bottom row*) with Perfect, Imperfect 1 and Imperfect 2 ranking

For each concomitant variable, we considered the SRS, RNS($\boldsymbol{\rho}_i, \zeta^*_{i,0.75}$) and RNS($\boldsymbol{\rho}_i, \zeta^{\mathrm{o}}_{i,0.75}$) designs where $\zeta^*_{i,0.75}$ is the ζ^* given by Theorem 3.9 for estimating $Q_{X,0.75}$ using $\boldsymbol{\rho}_i$ and $\zeta^{\mathrm{o}}_{i,0.75}$ is the optimum zeta for estimating $Q_{X,0.75}$ using $\boldsymbol{\rho}_i$, chosen as in Section 3.3. For each of these designs and each $p \in \{(112 + 7n)/224 : n = 0, \ldots, 16\}$, 5,000 confidence interval were obtained. For each interval $[Y_{r:m}, Y_{m-r+1}]$, the length was determined and whether $Q_{X,p}$ was in the interval.

Note that, to study the robustness of the RNS design to the choice of ζ, we calculated the best values of ζ designed for estimating $Q_{X,0.75}$ and used them throughout this section to construct confidence intervals for all quantiles $Q_{X,p}$ with $p \geq 0.5$. If the goal is to construct confidence intervals for an specific quantile $Q_{X,p}$, one should use an RNS design with suitable ζ obtained for $Q_{X,p}$ which results in better RNS confidence intervals than those obtained here.

Figure 6 shows the average confidence interval lengths. The top row contains the average lengths for the SRS and RNS($\boldsymbol{\rho}_i, \zeta^*_{i,0.75}$) designs for

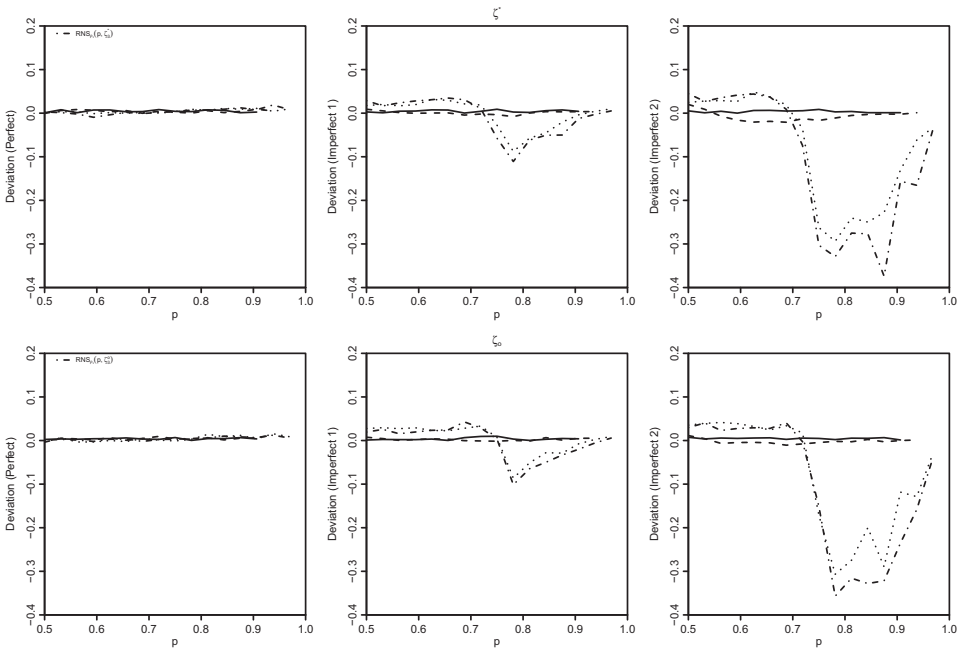

Figure 7: Average deviation from the exact coverage probabilities for 5,000 confidence intervals using $\zeta^*_{i,75}$ (*top row*) and $\zeta^{\mathrm{o}}_{i,0.75}$ (*bottom row*) with Perfect, Imperfect 1 and Imperfect 2 ranking

perfect ranking, imperfect ranking 1 (using the mothers weight at birth) and imperfect ranking 2 (the predicted values of regressing the mothers weight at birth and the lambs weight at birth on the seven-month weight). The second row contains the results for the RNS($\boldsymbol{\rho}_i, \zeta^{\circ}_{i,0.75}$) designs. Figure 7 shows the difference between the simulated coverage probability and the expected coverage probability. Of note is the fact that the designs based on $\boldsymbol{\rho}_2$ and $\boldsymbol{\rho}_3$ have very poor performance in terms of coverage probabilities when Kendal's τ is not large. The performance for the RNS($\boldsymbol{\rho}_1, \zeta^{*}_{1,0.75}$) and RNS($\boldsymbol{\rho}_1, \zeta^{\circ}_{1,0.75}$) designs is quite well, even under imperfect ranking. The reason for this is that the RNS($\boldsymbol{\rho}_1, \cdot$) is very close to simple random sampling. In general, as in Jafari Jozani and Johnson (2012), we recommend using a $\boldsymbol{\rho}$ which has large $\rho_1 = \mathbb{P}(K = 1)$ (say $> .75$) unless one is very confident about the ranking procedure.

5 Concluding Remarks

Randomized nomination sampling (RNS) was introduced by Jafari Jozani and Johnson (2012) and it has been shown to perform better than simple random sampling (SRS) for estimating the total and quantiles of the variable of interest in finite populations. RNS has a wide range of applications in environmental and ecological studies as well as medical research. In this paper, we obtained nonparametric confidence intervals for quantiles based on RNS data from continuous distributions. Several theoretical results regarding the distributional properties of RNS data as well as nonparametric confidence intervals for quantiles are obtained. We observed that the design parameter ζ associated with RNS technique is a very powerful and flexible tool to construct RNS designs resulting in confidence intervals with the desired coverage probabilities for a wide range of population quantiles without baring the use of randomized procedures. Numerical studies showed that, when the ranking process is perfect, the proposed RNS confidence intervals provide higher coverage probabilities and their expected length, especially for lower and upper quantiles, could be substantially shorter than their counterparts under SRS design. However, we observed that under imperfect ranking assumptions the average coverage probabilities of RNS confidence intervals could be very different from the desired nominal value. The results of the paper are consistent with the recommendation made by Jafari Jozani and Johnson (2012) regarding the choice of $\boldsymbol{\rho}$, the distribution on K. We observe that when $\boldsymbol{\rho}$ is such that $\rho_1 = \mathbb{P}(K = 1)$ is large, then RNS confidence intervals perform better, in terms of the average length, than their SRS counterparts under both perfect and imperfect ranking situations. However, under nearly

perfect ranking situations, one can obtain significantly better RNS confidence intervals especially for lower or upper quantiles compared with SRS design.

Acknowledgment. Authors gratefully acknowledge the research supports of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

BOYLES, R.A. and SAMANIEGO, F.J. (1986). Estimating a distribution function based on nomination sampling. *J. Am. Stat. Assoc.* **81**, 101–133.

BURGETTE, L.F. and REINTER, J.P. (2012). Modeling adverse birth outcomes via confirmatory factor quantile regression. *Biometrics* **68**, 92–100.

DAVID, H.A. and NAGARAJA, H.N. (2003) *Order statistics*, 3rd edn. Wiley.

JAFARI JOZANI, M. and JOHNSON, B.C. (2012). Randomized nomination sampling for finite populations. *J. Stat. Plann. Infer.* **142**, 2103–2115.

JAFARI JOZANI, M. and MIRKAMALI, S.J. (2011). Control charts for attributes with maxima nominated samples. *J. Stat. Plann. Infer.* **141**, 2386–2398.

JAFARI JOZANI, M. and MIRKAMALI, S.J. (2010). Improved attribute acceptance sampling plans based on maxima nomination sampling. *J. Stat. Plann. Infer.* **140**, 2448–2460.

KVAM, P.H. (2003). Ranked set sampling based on binary water quality data with covariates. *J. Agric. Biol. Environ. Stat.* **8**, 271–279.

KVAM, P.H. and SAMANIEGO, F.J. (1993). On estimating distribution functions using nomination samples. *J. Am. Stat. Assoc.* **88**, 1317–1322.

MURFF, E.J.T. and SAGER, T.W. (2006). The relative efficiency of ranked set sampling in ordinary least squares regression. *Environ. Ecol. Stat.* **13**, 41–51.

OZTURK, O., BILGIN, O. and WOLFE, D.A. (2005). Estimation of population mean and variance in flock management: a ranked set sampling approach in a finite population setting. *J. Stat. Comput. Simul.* **11**, 905–919.

NOURMOHAMMADI, M., JAFARI JOZANI, M. and JOHNSON, B. (2014). Confidence interval for quantiles in finite populations with randomized nomination sampling. *Comput. Stat. Data Anal.* **73**, 112–128.

TIWARI, R.C. (1988). Bayes estimation of a distribution under a nomination sampling. *IEEE Trans. Reliab.* **37**, 558–561.

TIWARI, R.C. and WELLS, M.T. (1989). Quantile estimation based on nomination sampling. *IEEE Trans. Reliab.* **38**, 612–614.

YU, P.L. and LAM, K. (1997). Regression estimator in ranked set sampling. *Biometrics* **53**, 1070–1080.

WELLS, M.T. and TIWARI, R.C. (1990) Estimating a distribution function based on minima-nomination sampling. In *Topics in statistical dependence, volume 16 of IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Hayward, pp. 471–479.

WILLEMAIN, T.R. (1980). Estimating the population median by nomination sampling. *J. Am. Stat. Assoc.* **75**, 908–911.

Appendix

Proof of Lemma 3.5

Since $H(\nu)$ is symmetric about $\nu = \frac{1}{2}$, it is enough to show that $H(\nu)$ is increasing for $\nu \in [0, \frac{1}{2}]$. To this end, by taking the first derivative of $H(\nu)$ with respect to ν, and after some manipulations, we have

$$\begin{aligned}\frac{d}{d\nu}H(\nu) &= \sum_{j=r}^{m-r} j\binom{m}{j}\nu^{j-1}(1-\nu)^{m-j} - \sum_{j=r}^{m-r}(m-j)\binom{m}{j}p^j(1-p)^{m-j-1} \\ &= m\binom{m-1}{r-1}\left\{nu^{r-1}(1-\nu)^{m-r} - \nu^{m-r}(1-\nu)^{r-1}\right\}.\end{aligned}$$

Now, $\frac{d}{d\nu}H(\nu) > 0 \iff \nu^{r-1}(1-\nu)^{m-r} > \nu^{m-r}(1-\nu)^{r-1}$, or equivalently $\nu^{m-2r+1} < (1-\nu)^{m-2r+1}$ or $\nu < \frac{1}{2}$ as $m - 2r > 0$ (since $r < m - r$).

Proof of Lemma 3.7

For any $\epsilon > 0$, we have

$$\mathbb{P}\left(|Y(\zeta) - Q_{X,1/2}| > \epsilon\right) = 1 - \left\{\mathbb{P}\left(Y(\zeta) \le Q_{X,1/2} + \epsilon\right) - \mathbb{P}\left(Y(\zeta) \le Q_{X,1/2} - \epsilon\right)\right\}.$$

When $\epsilon \ge \inf\{x; F(x) \ge 1\} - Q_{X,1/2}$ we have $\mathbb{P}\left(|Y(\zeta) - Q_{X,1/2}| > \epsilon\right) = \mathbb{P}\left(|X - Q_{X,1/2}| > \epsilon\right) = 0$. Let $\epsilon < \inf\{x; F(x) \ge 1\} - Q_{X,1/2}$. Then, we have

$$\begin{aligned}\mathbb{P}(|Y - Q_{X,1/2}| > \epsilon) = 1 - \Bigg\{ & \zeta \sum_{K=1}^{N} \rho_K F^k(Q_{X,1/2} + \epsilon) - (1-\zeta) \\ & \times \sum_{K=1}^{N} \rho_K (1 - \bar{F}^K(Q_{X,1/2} + \epsilon)) - \zeta \\ & \times \sum_{K=1}^{N} F^K(Q_{X,1/2} - \epsilon) - (1-\zeta) \\ & \times \sum_{K=1}^{N} (1 - \bar{F}^K(Q_{X,1/2} - \epsilon)) \Bigg\}.\end{aligned}$$

Since the population is symmetric about the median, we have $F(Q_{X,1/2}+\epsilon) = \bar{F}(Q_{X,1/2} - \epsilon)$ and $F(Q_{X,1/2} - \epsilon) = \bar{F}(Q_{X,1/2} + \epsilon)$. So

$$\mathbb{P}\left(|Y(\zeta) - Q_{X,1/2}| > \epsilon\right) = 1 - \left\{\sum_{K=1}^{N} \rho_K \left(F^K(Q_{X,1/2} + \epsilon) - F^K(Q_{X,1/2} - \epsilon)\right)\right\}.$$

Also, $F^K(Q_{X,1/2}+\epsilon) - F^K(Q_{X,1/2}-\epsilon)$ is a non-increasing function of K and so

$$\begin{aligned}\mathbb{P}\left(|Y(\zeta) - Q_{X,1/2}| > \epsilon\right) &\geq 1 - (F(Q_{X,1/2}+\epsilon) - F(Q_{X,1/2}-\epsilon))) \\ &= \mathbb{P}\left(|X - Q_{X,1/2}| > \epsilon\right)\end{aligned}$$

So, X is more peaked about $Q_{X,1/2}$ than $Y(\zeta)$ and this completes the proof.

Proof of Lemma 3.8

Let

$$\psi_\rho(p) = \frac{\beta_{\boldsymbol{\rho}}(p) - \frac{1}{2}}{\beta_{\boldsymbol{\rho}}(p) - \alpha_{\boldsymbol{\rho}}(p)}, \quad 0 < p < 1.$$

It is easy to show that

$$\psi_\rho\left(\frac{1}{2}+t\right) = 1 - \psi_{\boldsymbol{\rho}}\left(\frac{1}{2}-t\right), \quad 0 \leq t \leq \frac{1}{2}. \tag{1}$$

Since $\frac{\alpha_{\boldsymbol{\rho}}(p)}{\beta_{\boldsymbol{\rho}}(p)}$ is increasing in p ($p \geq 1/2$) and $\alpha_{\boldsymbol{\rho}}(p) \leq p \leq \beta_{\boldsymbol{\rho}}(p)$ we observe that $\frac{1}{1-\frac{\alpha_{\boldsymbol{\rho}}(p)}{\beta_{\boldsymbol{\rho}}(p)}}$ is non-decreasing in p. Now, since

$$\psi_{\boldsymbol{\rho}}(p) = \frac{1 - \frac{1}{2\beta_{\boldsymbol{\rho}}(p)}}{1 - \frac{\alpha_{\boldsymbol{\rho}}(p)}{\beta_{\boldsymbol{\rho}}(p)}},$$

and $\beta_{\boldsymbol{\rho}}(p)$ is non-decreasing in p and $\beta_{\boldsymbol{\rho}}(p) \geq p$, we observe that $1 - \frac{1}{2\beta_{\boldsymbol{\rho}}(p)}$ is non-decreasing and positive for all $p \geq \frac{1}{2}$. So, for all $p \geq \frac{1}{2}$, $\psi_{\boldsymbol{\rho}}(p)$ is non-decreasing in p. Using (1) it can be easily seen that $\psi_{\boldsymbol{\rho}}(p)$ is also non-decreasing in p for $p \leq \frac{1}{2}$ and this competes the proof.

Mohammad Nourmohammadi
Mohammad Jafari Jozani
Brad C. Johnson
Department of Statistics
University of Manitoba
Winnipeg, MB, R3T 2N2 Canada
E-mail: m_jafari_jozani@umanitoba.ca

Paper received: 10 July 2013; revised: 8 May 2014.