Ann Inst Stat Math (2009) 61:235–249
DOI 10.1007/s10463-007-0141-5

Exact two-sample nonparametric test for quantile difference between two populations based on ranked set samples

Omer Ozturk · N. Balakrishnan

Received: 4 September 2006 / Revised: 22 December 2006 / Published online: 10 August 2007

Abstract Creation of a ranked set sample, by its nature, involves judgment ranking error within set units. This ranking error usually distorts statistical inference of the population characteristics. Tests may have inflated sizes, confidence intervals may have incorrect coverage probabilities, and the estimators may become biased. In this paper, we develop an exact two-sample nonparametric test for quantile shift between two populations based on ranked set samples. This test is based on two independent exact confidence intervals for the quantile of interest corresponding to the two populations and rejects the null hypothesis of equal quantiles if these intervals are disjoint. It is shown that a pair of 83 and 93% confidence intervals provide a 5 and 1% test for the equality of quantiles. The proposed test is calibrated for the effect of judgment ranking error so that the test has the correct size even under a wide range of judgment ranking errors. A small scale simulation study suggests that the test performs quite well for cycle sizes as small as 2.

Keywords Sampling design · Sign test · Ranked set sampling · Judgment ranking · Median · Imperfect ranking · Calibration

1 Introduction

Let F and G be two continuous distribution functions with the quantiles of order p as $\xi_p = \inf\{x : F(x) \geq p\}$ and $\eta_p = \inf\{x : G(x) \geq p\}$, respectively. Exact

O. Ozturk (✉)
Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA
e-mail: omer@stat.ohio-state.edu

N. Balakrishnan
Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S 4K1
e-mail: bala@mc.master.ca

nonparametric confidence intervals for these quantiles can be constructed from the order statistics of simple random samples without any further conditions on F and G. These confidence intervals are available in the literature; see, for example, the books by Gibbons (1985), Rice (1995) and Arnold et al. (1992). Even though these confidence intervals are completely distribution-free, they have poor efficiency and so are not widely used in practice. The efficiency can be improved by using a sampling procedure that puts more structure in data. One such procedure is the ranked set sampling.

Ranked set sampling methodology creates artificial strata in data so that homogeneous observations are grouped together. It replaces independent and identically distributed observations $X_i, i = 1, \ldots, N$, with independent order statistics. To collect a ranked set sample, one selects Nk units from an infinite population and randomly divides these into N sets, each of size k. In each set, units are ranked without full measurement from smallest to largest. The i-th judgment ranked unit is fully measured in n_i sets of N so that $\sum_{i=1}^{k} n_i = N$. The fully measured observations $X_{[i]j}, j = 1, \ldots, n_i; i = 1, \ldots, k$, are called a ranked set sample. We note that $X_{[i]j}$ are independent but not identically distributed. If the quality of judgment ranking is perfect, the ranked set sample consists of independent order statistics and we use standard order statistic notation (replace square brackets with round one) to denote them by $X_{(i)j}, j = 1, \ldots, n_i; i = 1, \ldots, k$.

Even though the concept of ranked set sampling was introduced almost half a century ago by McIntyre (1952), literature in ranked set sampling has expanded rapidly only in the last two decades. Due to recent research interest in this area, the original paper of McIntyre on ranked set sampling was republished in McIntyre (2005). Nonparametric inference, in particular, has received considerable attention in the ranked set sampling methodology. Chen (2000) used independent order statistics in ranked set sample to construct confidence intervals for the population quantiles. Recently, Balakrishnan and Li (2006) and Ozturk and Deshpande (2006) used order statistics of a ranked set sample to construct exact nonparametric confidence intervals for population quantiles. These authors showed that the quantile intervals based on order statistics of a ranked set sample are narrower than those constructed from independent order statistics in Chen (2000). Deshpande et al. (2006) constructed quantile confidence intervals from a ranked set sample in the finite population context. Rank-based two-sample inference has also drawn considerable attention in the literature. Bohn and Wolfe (1992, 1994) introduced two-sample ranked set sample rank-sum test for the location shift between two populations. Further research on the two-sample problem are due to Ozturk (1999), and Ozturk and Wolfe (2000, 2001).

Most of these cited research, whether asymptotic or exact, relies on the assumption of perfect ranking. In absence of this assumption, tests may not have the correct size, confidence intervals may not achieve the desired coverage probability, and the estimators may not be unbiased. In this case, it is desirable to develop procedures that are efficient when we have perfect ranking and valid when we have some ranking error. Purpose of this paper is to achieve these two goals while making inference on a location shift between two populations. To achieve this goal, we need to have a valid imperfect ranking model.

Any sensible ranking model should be practical and flexible enough to cover a wide range of situations. There are mainly three classes of imperfect ranking models. The first model uses the mixture distribution of the actual order statistics to rank the observations within a set; see Bohn and Wolfe (1994) and Frey (2005). The second model uses an additive noise model due to David and Levine (1972) and Dell and Clutter (1972). In this model, units are ranked based on their perceived values that are tied to unmeasured values through an added noise variable. Finally, the third model uses the monotone likelihood ratio principle of Fligner and MacEachern (2006). In this work, ranked set sample observations are generated from Dell and Clutter (1972) model. In order to generate a ranked set sample from this imperfect ranking model, we generate two random vectors $\boldsymbol{u}$ and $\boldsymbol{w}$ each of size k, where u_i $(i = 1, \ldots, k)$ are i.i.d. from F and w_i $(i = 1, \ldots, k)$ are i.i.d. from a normal distribution with mean zero and variance σ_w^2. These vectors are added, $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{w}$, and the components of $\boldsymbol{v}$ are ordered to obtain an ordered set of $(v_{(i)}, u_{[i]})$. In this set, the value in the second component is taken as judgment ranked order statistics. In this process, for a given probability model F, the quality of judgment ranking is controlled by the noise variable w. If w is degenerate, then $u_{[i]} = v_{(i)}$; otherwise, ranking process will contain some error and the magnitude of this error depends on σ_w^2. For example, the correlation coefficient (ρ) between u and v is given by $\rho = 1/\sqrt{\sigma_F^2 + \sigma_w^2}$, where σ_F^2 is the variance of the underlying probability model F. If F is standard normal, then the selection of $\sigma_w^2 = 0, 7/9, 3$ produces correlation coefficient $\rho = 1, 0.75, 0.5$, respectively.

Our goal in this paper is to construct an exact two-sample nonparametric test for the difference between two population quantiles of order p. The proposed test uses individual confidence intervals of the population quantiles and rejects the null hypothesis if the intervals are disjoint. Section 2 introduces exact quantile intervals for a single population and discusses their properties. In order to achieve the desired coverage probability, quantile confidence intervals are constructed by interpolating the adjacent order statistics. Section 3 uses these interpolated confidence intervals to construct a two-sample test for quantile shift between the two populations. The test rejects the null hypothesis of equal population quantiles when the two interpolated confidence intervals are disjoint. Section 4 shows that the proposed test is robust against judgment ranking error and preserves the nominal type I error rate for cycle sizes as small as 2. Section 5 illustrates the test procedure by applying it to a two-sample ranked set data. Finally, Sect. 6 provides some concluding remarks.

2 One-sample quantile intervals

Let $X_{(i)j}$, $j = 1, \ldots, n_i$; $i = 1, \ldots, k$, be a ranked set sample from distribution F. The pdf and cdf of the ith order statistic from a simple random sample of size k are given by

$$f_{(r)}(x) = k\binom{k-1}{r-1} F^{r-1}(x)\{1 - F(x)\}^{k-r} f(x),$$

$$F_{(r)}(x) = \sum_{j=r}^{k} \binom{k}{j} F^{j}(x)\{1 - F(x)\}^{k-j}.$$

Let $X_{(1:N)} < \cdots < X_{(N:N)}$ be the ordered ranked set sample. We now seek a similar expression for the r-th order statistic of the ranked set sample. The ranked set sample observations are independent but not identically distributed. The derivation of the pdf and cdf of the rth order statistic in this case is not trivial and are given by (see Ozturk and Deshpande 2006; Balakrishnan and Li 2006)

$$f_{(r:N)}(x) = \sum_{i=1}^{k} n_i \sum_{U_{k,r-1,i}} \binom{n_i-1}{u_i-1} \prod_{j\neq i}^{k} \binom{n_j}{u_j} F_{(j)}^{u_j}(x)\left(1-F_{(j)}(x)\right)^{n_j-u_j} \times F_{(i)}^{u_i-1}(x)\left(1-F_{(i)}(x)\right)^{n_i-u_i} f_{(i)}(x),$$

where

$$U_{k,a,i} = \left\{(u_1,\ldots,u_k) : \sum_{j=1}^{k} u_j = a, 0 \le u_j \le n_j, 0 \le u_i \le n_i - 1, j \neq i\right\},$$

and

$$P(X_{(r:N)} \le y) = F_{r:N}(y) = \sum_{v=r}^{N} \sum_{U^*_{k,v}} \prod_{i=1}^{k} \binom{n_i}{u_i^*} \times (B_{i,k+1-i}(F(y)))^{u_i^*}(1-B_{i,k+1-i}(F(y)))^{n_i-u_i^*}, \quad (1)$$

where $B_{a,b}(y)$ is the cdf of the beta distribution with parameters (a, b) and

$$U^*_{k,v} = \left\{(u_1,\ldots,u_k) : \sum_{i=1}^{k} u_i^* = v, 0 \le u_i^* \le n_i\right\}.$$

In order to construct a $100(1-\alpha)\%$ confidence interval for ξ_p, we choose two integers r and s such that $1 \le s < r \le N$. The integers r, s can be determined by the following equality:

$$P(X_{(s:N)} \le \xi_p \le X_{(r:N)}) = 1 - \alpha.$$

This relationship may not determine r, s uniquely. In order to define r, s uniquely, we use equal tail probabilities so that r, s are uniquely defined by

$$\alpha/2 = P(X_{(s:N)} \ge \xi_p) = \sum_{v=0}^{s-1} \sum_{U^*_{k,v}} \prod_{i=1}^{k} \binom{n_i}{u_i^*} \times (B_{i,k+1-i}(F(\xi_p)))^{u_i^*}(1-B_{i,k+1-i}(F(\xi_p)))^{n_i-u_i^*} \quad (2)$$

and

$$\alpha/2 = P(X_{(r:N)} \leq \xi_p) = \sum_{v=r}^{N} \sum_{U^*_{k,v}} \prod_{i=1}^{k} \binom{n_i}{u^*_i} \times (B_{i,k+1-i}(F(\xi_p)))^{u^*_i} (1 - B_{i,k+1-i}(F(\xi_p)))^{n_i - u^*_i}, \tag{3}$$

where $N = \sum_{i=1}^{k} n_i$. This selection may not produce an interval that has exact coverage probability of $1-\alpha$ due to the discrete nature of the distribution. In order to approximate the coverage probability, Ozturk and Deshpande (2006) constructed confidence interval by interpolating adjacent order statistics in an ordered ranked set sample. By adapting their notation here, we consider two confidence intervals, $I_{s,r} = [X_{(s:N)}, X_{(r:N)}]$ and $II_{s+1,r-1} = [X_{(s+1:N)}, X_{(r-1:N)}]$, where confidence intervals $I_{s,r}$ and $II_{s+1,r-1}$ have confidence levels $1-\alpha_I$ and $1-\alpha_{II}$, $1-\alpha_{II} < 1-\alpha < 1-\alpha_I$. Then the interpolated confidence interval is constructed as

$$I_{\epsilon_1,\epsilon_3} = [L_s, U_r] = [(1-\epsilon_1)X_{(s:N)} + \epsilon_1 X_{(s+1:N)}, (1-\epsilon_2)X_{(r:N)} + \epsilon_2 X_{(r-1:N)}]$$

where

$$\epsilon_1 = \left[1 + \frac{s\{1-p\}(\alpha_{II}/2 - \alpha/2)}{(N-s)p(\alpha/2 - \alpha_I/2)}\right]^{-1}$$

and

$$\epsilon_2 = \left[1 + \frac{(N-(r-1))p(\alpha/2 - \alpha_{II}/2)}{(r-1)\{1-p\}(\alpha_I/2 - \alpha/2)}\right]^{-1}.$$

This interval provides a coverage probability close to the nominal value of $1-\alpha$. An exact confidence interval for η_p can be constructed in a similar fashion.

3 Two-sample exact test

Let $X_{(i)j}$, $j = 1, \ldots, n_i$; $i = 1, \ldots, k$, $N = \sum_{i=1}^{k} n_i$ and $Y_{(i)j}$, $j = 1, \ldots, m_i$; $i = 1, \ldots, k$, $M = \sum_{i=1}^{k} m_i$ be ranked set samples from distributions F and G, respectively. We now propose an exact test for hypothesis testing problem

$$H_0 : \Delta_p = \xi_p - \eta_p = 0 \quad \text{vs.} \quad H_A : \Delta_p \neq 0.$$

Let $I^x_{s,r} = [L_x, U_x] = [X_{(s:N)}, X_{(r:N)}]$ and $I^y_{s',r'} = [L_y, U_y] = [Y_{(s':M)}, Y_{(r':M)}]$ be the two confidence intervals for ξ_p and η_p with confidence coefficients $1-\alpha_x$ and $1-\alpha_y$, respectively. We reject the null hypothesis if the intervals $I^x_{s,r}$ and $I^y_{s',r'}$ are disjoint. Note that this testing procedure is distribution-free since the individual confidence intervals do not require any distributional assumptions. On the other hand, the individual confidence coefficients need to be determined to achieve the overall Type I error rate α_O.

The overall Type I error rate is the probability of rejecting the null hypothesis when it is true. In this case, we reject H_0 if the confidence intervals are disjoint. This is possible when either $U_x < L_y$ or $U_y < L_x$. The Type I error is thus given by

$$\begin{aligned}\alpha_O &= P_{H_0}(U_x < L_y) + P_{H_0}(U_y < L_x)\\ &= P_{H_0}(X_{(r:N)} < Y_{(s':M)}) + P_{H_0}(Y_{(r':M)} < X_{(s:N)}).\end{aligned}$$

Theorem 1 *Under perfect ranking, for given r, s, r', s', the Type I error is given by*

$$\alpha_O = \int_0^1 F_{(r:N)}(v)\mathrm{d}F_{(s':M)}(v) + \int_0^1 F_{(r':M)}(v)\mathrm{d}F_{(s:N)}(v),$$

where

$$\begin{aligned}\int_0^1 F_{(r:N)}(v)\mathrm{d}F_{(s':M)}(v) = \sum_{v=r}^{N}\sum_{U^*_{k,v}}\sum_{t=1}^{k}\sum_{U_{k,s'-1,t}} \int \Bigg\{ &\prod_{i=1}^{k}\binom{n_i}{u_i^*}(B_{i,k+1-i}(v))^{u_i^*}\\ &\times(1-B_{i,k+1-i}(v))^{n_i-u_i^*}\\ &\times m_t\binom{m_t-1}{u_t-1}B^{u_t}_{t,q+1-t}(v)\big(1-B_{t,q+1-t}(v)\big)^{m_t-u_t}b_{t,q+1-t}(v)\\ &\times\prod_{j\neq t}^{k}\binom{m_j}{u_j}B^{u_j-1}_{j,k+1-j}(v)\left(1-B_{j,k+1-j}(v)\right)^{m_j-u_j}\mathrm{d}v\Bigg\}\end{aligned}$$

and

$$\begin{aligned}\int_0^1 F_{(r':M)}(v)\mathrm{d}F_{(s:N)}(v) = \sum_{v=r'}^{M}\sum_{U^*_{q,v}}\sum_{t=1}^{q}\sum_{U_{q,s'-1,t}} \int \Bigg\{ &\prod_{i=1}^{q}\binom{m_i}{u_i^*}(B_{i,q+1-i}(v))^{u_i^*}\\ &\times(1-B_{i,q+1-i}(v))^{m_i-u_i^*}\\ &\times n_t\binom{n_t-1}{u_t-1}B^{u_t}_{t,k+1-t}(v)\\ &\times\left(1-B_{t,k+1-t}(v)\right)^{n_t-u_t}b_{t,k+1-t}(v)\\ &\times\prod_{j\neq t}^{k}\binom{n_j}{u_j}B^{u_j-1}_{j,k+1-j}(v)\left(1-B_{j,k+1-j}(v)\right)^{n_j-u_j}\mathrm{d}v\Bigg\}.\end{aligned}$$

Note that the integrals in the above theorem do not involve any unknown quantity revealing that the test is still distribution-free. Even though it is possible to expand the integral into series of sums, it is easier and convenient to perform the necessary numerical integration to compute the Type I error.

In practice, to perform the test, one needs to determine two confidence coefficients for the confidence intervals of ξ_p and η_p so that the desired Type I error is achieved. The selection of these confidence intervals (or equivalently selection of s, r, s', r') is not unique. One way to specify these quantities uniquely is to select confidence intervals with equal confidence coefficients each having equal tail probabilities. In order to achieve this, we propose the following algorithm:

I Set α_O and then construct confidence intervals for ξ_p and η_p with confidence coefficients $1-\alpha_0$ each having equal tail probabilities, i.e., determine r, s, r', s'.

II By using Theorem 1, compute the new Type I error, α_N, for r, s, r', s'.

III If $\alpha_N < \alpha_O$, then update confidence intervals $r = r+1, s = s-1, r' = r'+1, s' = s'-1$ and go to step II. If $\alpha_N \approx \alpha_O$, then use the current r, s, r', s' to construct the test.

For finite samples, due to the discrete nature of the sign statistics, the desired Type I error rate α_O may not be available from this algorithm. In this case, we may select largest r, r' and smallest s, s' so that $\alpha_N \leq \alpha_O$.

Ozturk (1999) looked at a similar testing procedure for large sample sizes. He showed that for a 5 and 1% two-sample test, one needs to construct a pair of roughly 83 and 93% confidence intervals. For small cycle sizes, by using the above algorithm, we also show that 83 and 93% confidence intervals provide roughly 5 and 1% two-sample median tests.

Table 1 presents coverage probabilities of the individual confidence intervals and Type I error rates of the proposed two-sample test for perfect and imperfect ranking without any calibration. All entries in perfect ranking ($\rho = 1$), except $I_{\epsilon_1,\epsilon_2}$, are exact and computed from the above algorithm. Entries for imperfect ranking and $I_{\epsilon_1,\epsilon_2}$ are obtained from a small scale simulation study in which the ranked set samples were generated from the imperfect judgment ranking model of Dell and Clutter (1972) and David and Levine (1972). Underlying distribution F was taken as standard normal. The perfect judgment ranked set samples were generated from the Dell and Clutter model with $\rho = 1$ and imperfect judgment ranked set samples were generated with $\rho = 0.50, 0.75$. Simulation size was taken as 2000.

It is obvious that, under perfect ranking, interpolated confidence intervals provide coverage probabilities very close to the nominal values (83%) for cycle sizes as small as 2. This close approximation also provides a Type I error rate α_O that is close to the nominal size (0.05) of the test. Similar results also hold for a 1% test with a pair of 93% confidence intervals, and for brevity the corresponding results are not reported here. It is important to note here that these coverage probabilities and Type I error rates are exact and do not rely on asymptotic theory.

Under imperfect ranking ($\rho = 0.75, 0.5$) with no calibration for the impact of ranking error, the interpolated confidence intervals provide coverage probabilities that are smaller than the nominal values. Consequently, the Type I error rates are inflated. This shows that there is a need to calibrate the effect of imperfect ranking on coverage probabilities and hence on the Type I error rates. The next section provides a procedure to calibrate the effect of judgment ranking error on the proposed two-sample test.

Table 1 Coverage probabilites of median ($p = 0.5$) confidence intervals and Type I error of the two sample test under perfect ($\rho = 1$) and imperfect ranking ($\rho = 0.75, 0.5$)

				$\rho = 1$		$\rho = 0.75$		$\rho = 0.5$	
p	Cycle	Set	CI	$1-\alpha$	α_O	$1-\alpha$	α_O	$1-\alpha$	α_O
0.5	2	2	$I_{1,4}$	0.930	0.010	0.908	0.013	0.891	0.024
			$I_{2,3}$	0.461	0.378	0.407	0.427	0.399	0.465
			$I_{\epsilon_1,\epsilon_2}$	0.863	0.039	0.823	0.048	0.795	0.075
0.5	3	2	$I_{2,5}$	0.855	0.038	0.817	0.065	0.789	0.069
			$I_{3,4}$	0.376	0.484	0.342	0.547	0.322	0.546
			$I_{\epsilon_1,\epsilon_2}$	0.832	0.052	0.799	0.073	0.770	0.089
0.5	4	2	$I_{2,7}$	0.965	0.003	0.947	0.009	0.940	0.009
			$I_{3,6}$	0.790	0.074	0.744	0.110	0.727	0.117
			$I_{\epsilon_1,\epsilon_2}$	0.828	0.061	0.786	0.086	0.770	0.087
0.5	5	2	$I_{3,8}$	0.938	0.008	0.916	0.014	0.897	0.020
			$I_{4,7}$	0.735	0.112	0.699	0.146	0.675	0.184
			$I_{\epsilon_1,\epsilon_2}$	0.829	0.049	0.805	0.065	0.775	0.094
0.5	2	3	$I_{2,3}$	0.900	0.020	0.831	0.047	0.807	0.076
			$I_{3,4}$	0.402	0.452	0.365	0.500	0.319	0.571
			$I_{\epsilon_1,\epsilon_2}$	0.864	0.032	0.780	0.074	0.747	0.110
0.5	3	3	$I_{3,7}$	0.919	0.014	0.866	0.040	0.835	0.042
			$I_{4,6}$	0.611	0.220	0.544	0.290	0.520	0.319
			$I_{\epsilon_1,\epsilon_2}$	0.859	0.039	0.792	0.077	0.745	0.101
0.5	4	3	$I_{4,9}$	0.939	0.008	0.894	0.018	0.876	0.031
			$I_{5,8}$	0.735	0.113	0.676	0.174	0.637	0.197
			$I_{\epsilon_1,\epsilon_2}$	0.851	0.045	0.785	0.087	0.746	0.109
0.5	5	3	$I_{5,11}$	0.954	0.004	0.908	0.017	0.891	0.019
			$I_{6,10}$	0.816	0.060	0.742	0.098	0.717	0.131
			$I_{\epsilon_1,\epsilon_2}$	0.841	0.051	0.757	0.090	0.734	0.118

4 Imperfect ranking

Practical application of ranked set sampling often involves imperfect ranking by the nature of sampling procedure. Thus, it is essential to study any inferential procedure based on ranked set sampling under imperfect ranking. Even though the proposed testing procedure is distribution-free under perfect ranking, this does not imply that it will still be distribution-free under any judgment ranking scheme. Under imperfect ranking, the cdf of the rth order statistic in Eq. (1) does not hold in general. The problem in this case is that under imperfect ranking the cdf of the ith judgment order statistic can not be written as an incomplete beta function as in the perfect ranking case. To resolve this problem, we estimate the cdf of the ith judgment order statistic and insert this estimate into Eqs. (2) and (3) to construct the confidence intervals. A natural estimator for the cdf of $F_{[i]}(\xi_p)$ would be the empirical cdf of the ith judgment class distribution $\hat{F}_{[i]}(\xi_p)$. On the other hand, for small cycle sizes, these estimators may have undesirable properties that violates the stochastic order relation of the order statistics, such as $\hat{F}_{[i]}(\xi_p) < \hat{F}_{[i+1]}(\xi_p)$. Recently, Ozturk (2007) developed an estimator for the cdf of the ith, $i = 1, \ldots, k$, judgment class distribution that does not violate this stochastic order restriction among the judgment classes, given by

$$F_{[i]N}(\xi_p) = \min_{1 \le r \le i} \max_{i \le s \le k} A_{r,s}(\xi_p), \tag{4}$$

where

$$A_{r,s}(\xi_p) = \frac{\sum_{u=r}^{s} n_i \hat{F}_{[i]}(\xi_p)}{\sum_{u=r}^{s} n_i},$$

and $\hat{F}_{[i]}(\xi_p)$ is the empirical cdf of the ith judgment order statistic at the pth quantile of the underlying distribution.

Ozturk (2007) showed that this estimator has smaller integrated mean square error (IMSE) and mean square error (MSE) at every point of the support of the distribution than the IMSE and MSE of the empirical cdf estimators.

To calibrate the effect of imperfect ranking, we compute the individual confidence intervals from Eqs. (2) and (3) by replacing $B_{i,k+1-i}(F(\xi_p))$ with an appropriate estimate. For simplicity, let us assume that $k = q$. In order to estimate ξ_p, we first center X- and Y-observations by subtracting their medians. We then estimate the pth quantile from the combined sample of centered X- and Y-observations. Let $\hat{\xi}_p$ be this estimate. Then, for the estimation of $F_{[i]}(\hat{\xi}_p)$, we combine the centered ith, $i = 1, \ldots, k$, judgment class observations, viz., $X_{[i]j} - \hat{\xi}_p (j = 1, \ldots, n_i)$ and $Y_{[i]j} - \hat{\xi}_p (j = 1, \ldots, m_i)$, from X- and Y-samples, to obtain $Z_{[i]j}$ $(j = 1, \ldots, n_i + m_i)$ and compute $\hat{F}_{[i]}(\hat{\xi}_p)$ from this combined sample. Then, the estimator $F_{[i]K}(\hat{\xi}_p)$ is obtained from Eq. (4), where $K = N + M$.

Even though $F_{[i]K}(\hat{\xi}_p)$ has nice properties, for small cycle and set sizes it is possible that it can be either zero or one which creates computational problem in Eqs. (2) and (3). To resolve this problem, we introduce a truncated version of $F_{[i]K}(\hat{\xi}_p)$. Note that for any sensible ranking procedure, $F_{[i]}(\xi_p)$ varies between the two extreme cases of perfect $(B_{i,k+1-i}(p))$ and random (p) rankings. Thus, it lies in the interval $[B_{i,k+1-i}(p), p]$ for $i = 1, \ldots, (k+1)/2$ and $[p, B_{i,k+1-i}(p)]$ for $i = (k+1)/2, \ldots, k$. By using this property, we introduce a truncated estimator. The truncated estimator, $\tilde{F}_{[i]K}(\hat{\xi}_p)$, takes the value of $F_{[i]K}(\hat{\xi}_p)$ if it is in the interval and the closest value otherwise:

$$\tilde{F}_{[i]K}(\hat{\xi}_p) = \begin{cases} F_{[i]K}(\hat{\xi}_p) & p \le F_{[i]K}(\hat{\xi}_p) \le B_{i,k+1-i}(p) \\ B_{i,k+1-i}(p) & F_{[i]K}(\hat{\xi}_p) > B_{i,k+1-i}(p) \\ p & F_{[i]K}(\hat{\xi}_p) < p \end{cases} \tag{5}$$

for $i = 1, \ldots, (k+1)/2$ and

$$\tilde{F}_{[i]K}(\hat{\xi}_p) = \begin{cases} F_{[i]K}(\hat{\xi}_p) & B_{i,k+1-i}(p) \leq F_{[i]K}(\hat{\xi}_p) \leq p \\ B_{i,k+1-i}(p) & F_{[i]K}(\hat{\xi}_p) < B_{i,k+1-i}(p) \\ p & F_{[i]K}(\hat{\xi}_p) > p \end{cases} \tag{6}$$

for $i = (k+1)/2, \ldots, k$.

For the implementation of the proposed two-sample test, we first estimate the pth quantile of the population and the CDF of the ith judgment class distributions at $\hat{\xi}_p$ for $i = 1, \ldots, k$. We then construct the individual interpolated confidence intervals for the specified Type I error rate. For example, for a 5% test we construct a pair of 83% interpolated confidence intervals by using Eqs. (2) and (3), where we replace $B_{i,k+1-i}(p)$ with $\tilde{F}_{[i]K}(\hat{\xi}_p)$, $i = 1, \ldots, k$. Finally, we reject the null hypothesis if these confidence intervals are disjoint.

In order to evaluate the performance of the proposed test procedure, we performed a simulation study. Ranked set samples in this simulation study were generated from the Dell and Clutter model with $\rho = 1, 0.75, 0.50$ and given set ($k = 2, 3$) and cycle ($n = 2(1)10$) size specifications. Each entry was computed from 2,000 Monte Carlo simulations.

Table 2 presents empirical coverage probabilities $(1-\alpha)$ of the interpolated confidence intervals and the empirical Type I error rates (α_O) of the two-sample test for both ranked set and simple random samples when the quantile of interest is the median ($p = 0.5$). For efficiency comparison, we also provide average length (AL) of the interpolated confidence intervals.

From Table 2, it is clear that the coverage probabilities and the Type I error rates are reasonably close to the nominal values across all ranking qualities for moderately small set and cycle sizes. For example, the coverage probabilities of the interpolated confidence intervals are mostly varying between 85 and 81% which yield Type I error rates around 5% for the test. For the poor ranking quality ($\rho = 0.5$) and smaller sample sizes, the coverage probabilities of the confidence intervals tend to be slightly smaller than 81% when the set sizes are odd. On the other hand, this phenomenon disappears when sample sizes get large. For example, when $\rho = 0.5$, and (n, k) are $(3, 3)$, $(5, 3)$, $(7, 3)$ and $(9, 3)$, the coverage probabilities are 0.791, 0.806, 0.816, 0.808, respectively. These values are smaller than the coverage probabilities for other (n, k) values, but there is an increasing trend in these coverage probabilities with increased sample sizes. We believe that the reason for this phenomenon is due to the way that the truncated estimators of the judgment class distributions are computed in Eqs (5) and (6). Even with these low values of coverage probabilities, the proposed test achieves its nominal size reasonably well. Similar results also hold at other quantiles such as at $p = 0.25, 0.75$, but they are not presented here for the sake of brevity.

The performance of the proposed test for median quantiles is also evaluated in terms of its empirical power curves in Fig. 1 when set and cycle sizes are 3 and 4, respectively. Again the ranked set samples were generated from the Dell and Clutter model with the correlation coefficient $\rho = 1, 0.75$ and 0.5. For comparison purposes, the power curve of simple random sample two-sample median test is also provided.

Table 2 Simulated coverage probabilities $(1-\alpha)$ of interpolated confidence intervals of population median and Type I error rates (α_O) of the two-sample test after adjustment for imperfect ranking

n	k	rss, $\rho = 1$			rss, $\rho = 0.75$			rss, $\rho = 0.5$			srs		
		$1-\alpha$	α_O	AL	$1-\alpha$	α_O	AL	$1-\alpha$	α_O	AL	$1-\alpha$	α_O	AL
2	3	0.870	0.030	1.209	0.825	0.047	1.227	0.802	0.066	1.269	0.824	0.052	1.382
3	3	0.825	0.032	0.947	0.806	0.053	0.989	0.788	0.059	1.020	0.829	0.049	1.161
4	3	0.843	0.037	0.809	0.814	0.057	0.852	0.813	0.061	0.908	0.827	0.048	0.998
5	3	0.841	0.041	0.736	0.809	0.059	0.825	0.806	0.057	0.834	0.824	0.061	0.866
6	3	0.846	0.034	0.671	0.815	0.056	0.711	0.813	0.065	0.748	0.833	0.056	0.794
7	3	0.825	0.043	0.614	0.813	0.054	0.657	0.816	0.052	0.695	0.823	0.053	0.735
8	3	0.858	0.043	0.577	0.826	0.052	0.612	0.821	0.058	0.649	0.830	0.051	0.693
9	3	0.826	0.053	0.547	0.818	0.053	0.581	0.808	0.053	0.623	0.820	0.044	0.653
10	3	0.835	0.041	0.514	0.836	0.049	0.560	0.830	0.049	0.588	0.830	0.063	0.613
2	2	0.863	0.029	1.542	0.831	0.044	1.555	0.820	0.055	1.564	0.840	0.042	1.725
3	2	0.854	0.040	1.303	0.823	0.057	1.300	0.824	0.064	1.321	0.820	0.060	1.366
4	2	0.838	0.044	1.077	0.824	0.057	1.088	0.815	0.067	1.107	0.826	0.048	1.167
5	2	0.854	0.034	0.961	0.831	0.046	0.988	0.828	0.056	1.019	0.827	0.051	1.062
6	2	0.849	0.043	0.879	0.836	0.049	0.913	0.825	0.061	0.935	0.829	0.057	0.987
7	2	0.856	0.037	0.843	0.825	0.051	0.858	0.827	0.049	0.875	0.825	0.049	0.924
8	2	0.845	0.047	0.773	0.834	0.052	0.801	0.834	0.052	0.828	0.832	0.049	0.850
9	2	0.830	0.040	0.729	0.828	0.056	0.750	0.829	0.057	0.768	0.828	0.057	0.792
10	2	0.847	0.046	0.688	0.833	0.049	0.714	0.832	0.054	0.727	0.831	0.049	0.755

The power curves indicate that the size of the test is reasonably close to nominal size 0.05. As expected, the high power of ranked set sample two-sample median test when compared to the simple random sample two-sample median test is confirmed in this simulation study even when the correlation is as low as $\rho = 0.5$.

5 An illustrative example

To illustrate the use of the proposed two-sample median test procedure, we apply the suggested procedure to a ranked set sample data. The data were generated from Helsel and Hirsch (2002). This data set contains measurements of uranium concentration (part per billion) and total dissolved solids (TDS) in ground waters in two different types of aquifers. The first type aquifer contains ground water with bicarbonate (HCO_3) concentration level of at most 50% while the second type aquifer contains ground water with bicarbonate (HCO_3) concentration level of more than 50%. Here, we would like to test if the median uranium levels in low and high bicarbonate aquifers are different. In this setting, precise measurement of uranium concentration is expensive when compared with the cost of ranking water specimens based on TDS, but these

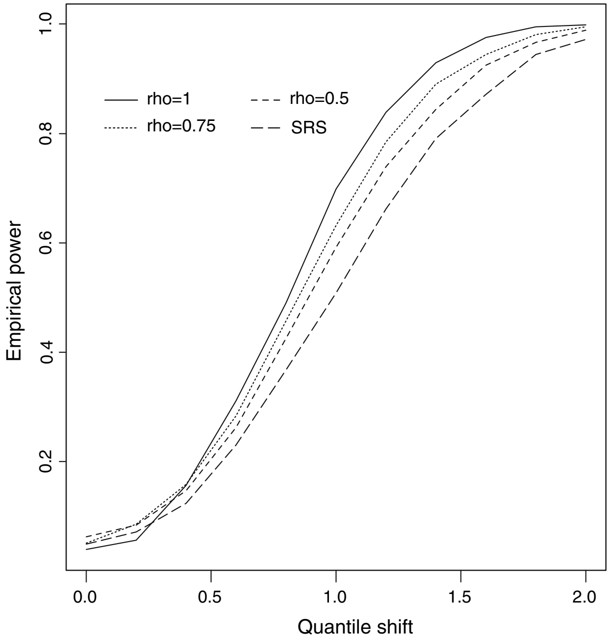

Fig. 1 Empirical power of two-sample median test when $k = 3$ and $n = 4$

Table 3 Ranked set sample of uranium concentrations (part per bilion) from low ($\leq 50\%$) and high ($> 50\%$) bicarbonate (HCO_3) aquifers

Rank	Low		High	
Cycle	1	2	1	2
1	0.9315	11.9042	4.7360	5.6290
2	1.5674	0.9772	3.0950	11.2724
3	0.4367	10.1142	4.9807	14.6342
4	0.4806	6.0876	1.5291	6.3042
5	0.1473	3.0918	0.9672	2.1568

two variables are highly correlated. Thus, it is appropriate to use ranked set sampling procedure.

Data set in Helsel and Hirsch (2002) contains 23 and 20 water specimens from low and high bicarbonate aquifers. The correlation coefficients between the uranium concentration and TDS are 0.64 and 0.84 for low and high bicarbonate aquifers, respectively. In order to create a two-sample ranked set sample, we select set size $k = 2$ and cycle size $n = 5$. For each sample, viz., low and high bicarbonate level, we randomly selected four water specimens and divided them into two sets, each of size 2. Units in each set is ranked based on TDS measurement. We then use the uranium concentration from the specimen that has the lowest TDS measurement in one set and the highest TDS measurement in the other. This process is repeated for five cycles to obtain ten uranium concentration for each sample. The ranked set sample thus obtained is given in Table 3.

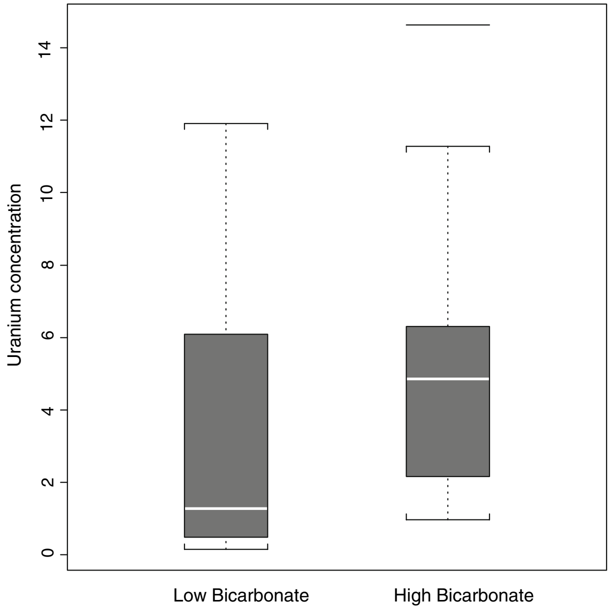

Fig. 2 Box plot of uranium concentration data

Figure 2 presents the box plots of the uranium concentrations for low and high bicarbonate ranked set samples. From a visual inspection of these box plots, it appears that low bicarbonate water wells contain slightly higher uranium concentrations than the uranium concentration of the high bicarbonate water wells.

Let μ_L and μ_H be the median uranium level of low and high uranium concentrations. We are interested in testing the hypothesis

$$H_0 : \mu_L - \mu_H = 0 \text{ against } H_0 : \mu_L - \mu_H \neq 0.$$

To perform the test, we first center the low- and high-bicarbonate samples by subtracting their medians 1.2723 and 4.85835, respectively, and compute $\hat{F}_{[i]20}(0)$ from the combined centered observations, $\hat{F}_{[1]20}(0) = 0.85$ and $\hat{F}_{[2]20}(0) = 0.20$. Under perfect ranking, $F_{(1)}(0) = 0.750$ and $F_{(2)}(0) = 0.250$. Thus, the truncated estimates of $F_{[1]}(0)$ and $F_{[2]}(0)$ from Eqs. (5) and (6) are $\tilde{F}_{[1]20}(0) = 0.750$ and $\tilde{F}_{[2]20}(0) = 0.250$. Let $X^*_{(1)} < \cdots < X^*_{(10)}$ and $Y^*_{(1)} < \cdots < Y^*_{(10)}$ be the ordered ranked set samples from low- and high-bicarbonate samples. The confidence intervals for these two samples are constructed as described in Sect. 2. These intervals based on X- and Y-samples are

$\{X^*_{(3)}, X^*_{(8)}\}$	$\{0.4806, 6.0876\}$	93.80%
$\{X^*_{(4)}, X^*_{(7)}\}$	$\{0.9315, 3.0919\}$	73.54%
$I^x_{0.821,0.500}$	$\{0.851, 4.590\}$	83.0%

and

$\{Y^*_{(3)}, Y^*_{(8)}\}$	$\{2.1568, 6.304\}$	93.80%
$\{Y^*_{(4)}, Y^*_{(7)}\}$	$\{3.095, 5.626\}$	73.54%
$I^y_{0.821,0.500}$	$\{2.926, 5.967\}$	83.0% .

We note that the 83% confidence intervals $I^x_{0.821,0.500}$ and $I^y_{0.821,0.500}$ are not disjoint. Therefore, we fail to reject the null hypothesis at 5% significance level and conclude that the median uranium concentration in low- and high-bicarbonate aquifers are not statistically different at 5% significance level.

6 Conclusions

In situations wherein recruiting a unit for a study is substantially cheaper than making formal measurements, procedures based on ranked set sampling provide substantial improvement in efficiency over procedures based on simple random sampling. This improved efficiency results from the additional structure provided by units that are used in judgment ranking process. Most of the theoretical results in ranked set sampling heavily depend on the very strong assumption of perfect ranking of this judgment ranking process. This assumption has a big impact on the efficiency as well as the validity of the inferential procedures. Thus, it is essential to develop inference based on ranked set samples that minimizes the effect of judgment ranking error. This problem is addressed in this research.

We have developed a two-sample nonparametric inference to test the equality of quantiles of two populations. The proposed test requires to construct a pair of 83 and 93% confidence intervals for a 5 and 1% tests. These confidence intervals are constructed by interpolating adjacent order statistics to achieve the desired coverage probabilities. Confidence intervals are also calibrated in order to minimize the impact of imperfect ranking. The simulation study shows that the proposed test maintains its nominal size for cycle sizes as small as 2. We have applied the proposed test to a ranked set sample data to illustrate how it could be used in practice.

An asymptomatic test that uses the similar idea of the present paper in the multi-sample situation has been discussed in the work of Ozturk and MacEachern (2004), but an exact test will be worth developing along the lines of the two-sample situation discussed in the present paper. However, its construction and properties need to be carefully examined and this will be the subject matter of a future paper.

References

Arnold, B. C., Balakrishnan, N., Nagaraja, H. N. (1992). *A first course in order statistics*. New York: Wiley.

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

Bohn, L. L., Wolfe, D. A. (1992). Nonparametric two-sample procedures for ranked-set sample data. *Journal of the American Statisttical Association, 87*, 552–562.

Bohn, L. L., Wolfe, D. A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann–Whitney–Wilcoxon statistics. *Journal of the American Statistical Association, 89*, 168–176.

Chen, Z. (2000). On ranked set sample quantiles and their applications. *Journal of Statistical Planning and Inference, 83*, 125–135.

David, H. A., Levine, D. N. (1972). Ranked set sampling in the presence of judgment error. *Biometrics, 28*, 553–555.
Dell, T. R., Clutter, J. L. (1972). Ranked-set sampling theory with order statistics background. *Biometrics, 28*, 545–555.
Deshpande J. V., Frey, J., Ozturk, O. (2006). Nonparametric ranked-set sampling confidence intervals for quantile of a finite population. *Environmental and Ecological Statistics, 13*, 25–40.
Fligner, M., MacEachern, S. N. (2006). Ranked set sampling: Models and distribution free two sample methods under imperfect ranking. *Journal of the American Statistical Association, 101*, 1107–1118.
Frey, J. (2005). Inference procedures based on order statistics. Ph.D. Thesis, Department of Statistics, The Ohio State University, Columbus, Ohio.
Gibbons, J. D. (1985). *Nonparametric statistical inference*. New York: Marcel Dekker.
Helsel, D. R., Hirsch, R. M. (2002). Statistical Methods in Water Resources, U. S. Geological Surveys, http:/water.usgs.gov/pubs/twri/twri4a3.
McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked-sets. *Australian Journal of Agricultural Research, 3*, 385–390.
McIntyre, G. A. (2005). A method of unbiased selective sampling using ranked sets. *The American Statistician, 3*, 230–232.
Ozturk, O. (1999). Two-sample inference based on one-sample ranked set sample sign statistics. *Journal of Nonparametric Statistics, 10*, 197–212.
Ozturk, O. (2007). Statistical inference under a stochastic ordering constraint in ranked set sampling. *Journal of Nonparametric Statistics* (in print).
Ozturk, O., Deshpande, J. V. (2006). Ranked-set sample nonparametric quantile confidence intervals. *Journal of Statistical Planning and Inference, 136*, 570–577.
Ozturk, O., MacEachern, S. (2004). Order restricted randomized designs for control versus treatment comparison. *Annals of Institute of Statistical Mathematics, 56*, 701–720.
Ozturk, O., Wolfe, D. A. (2000). Optimal allocation procedure in ranked set two-sample median test. *Journal of Nonparametric Statistics, 13*, 57–76.
Ozturk, O., Wolfe, D. A. (2001). An improved ranked set two-sample Mann–Whitney–Wilcoxon test. *Canadian Journal of Statistics, 28*, 123–135.
Rice, J. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury Press.