Ann. Inst. Statist. Math.
Vol. 56, No. 4, 701–720 (2004)

ORDER RESTRICTED RANDOMIZED DESIGNS FOR CONTROL VERSUS TREATMENT COMPARISON

OMER OZTURK AND STEVEN N. MACEACHERN

Department of Statistics, The Ohio State University, Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, U.S.A.

(Received May 2, 2003; revised December 18, 2003)

Abstract. We develop a method of randomizing units to treatments that relies on subjective judgement or on possible coarse modeling to produce restrictions on the randomization. The procedure thus fits within the general framework of ranked set sampling. However, instead of selecting a single unit from each set for full measurement, all units within a set are used. The units within a set are assigned to different treatments. Such an assignment translates the positive dependence among units within a set into a reduction in variation of contrasting features of the treatments.

A test for treatment versus control comparison, with controlled familywise error rate, is developed along with the associated confidence intervals. The new procedure is shown to be superior to corresponding procedures based on completely randomized or ranked set sample designs. The superiority appears both in asymptotic relative efficiency and in power for finite sample sizes. Importantly, this test does not rely on perfect rankings; rather, the information in the data on the quality of rankings is exploited to maintain the level of the test when rankings are imperfect. The asymptotic relative efficiency of the test is not affected by estimation of the quality of rankings, and the finite sample performance is only mildly affected.

Key words and phrases: Ranked set sampling, sampling design, sign test.

1. Introduction

In many applications, there are several potential measurements of a response that may be made, ranging from an expensive or time-consuming gold standard through quicker, numerical methods that have substantial bias or measurement error, to fully subjective forecasts that give only the rank order of a small set of the gold-standard measurements. These subjective forecasts may be accurate, or they may misorder the gold-standard measurements. A classic example of this application is the estimation of a pasture yield described by McIntyre (1952). In his motivating example, the field's yield is to be estimated by close clipping of a number of quadrats, each one meter square, creating an estimate of the yield per quadrat, μ, and thence the total yield of the pasture. The gold-standard measurement is the yield of each quadrat, measured by close clipping. Ignoring the cost of measurement, the best estimate of μ would be created by harvesting and measuring the yield of each quadrat; then averaging. However, the cost of measuring a quadrat's yield is substantial. A conventional design would choose a simple random sample (srs) of n quadrats and measure the yield of each of these quadrats. The mean of the measured quadrats, $\bar{Y}_{srs}$, provides an estimate of μ. However, the above strategy ignores substantial information that is available to the experimenter—namely subjective

estimation of the relative yield of various quadrats.

An alternative cheap design is to partition the pasture into $N \geq n$ sets, with each set consisting of, say, k quadrats. The ranker visually inspects each set of k quadrats and attempts to order them, from least yield to greatest yield. A simple random sample of n sets is chosen. Each selected set produces a single measured observation. To do this, the n selected sets are randomly partitioned into k groups of sets, each of size n/k. In the first group, the quadrat judged to have the least yield is measured; in the second group, the quadrat judged to have the second least yield is measured, and so on, until, in the k-th group, the quadrat judged to have the greatest yield is measured. Exactly one quadrat is measured from each set. A blinding technique is used to prevent hidden biases from entering the selection procedure. This design, called a balanced ranked set sample (rss), makes use of the ranker's subjective judgement to create k strata. A quadrat's stratum is the judged rank of the quadrat's yield within a set. Having created these strata, the sample is nearly a stratified random sample. The dependence of the samples across the strata (notice that two units in same set cannot both appear in the sample) is what makes this sampling plan different. However, when one passes from the finite population setting to an infinite population setting, where $N = \infty$ and the ranked sets are generated as i.i.d. draws from some distribution, the samples from the different strata are independent random samples from the within-stratum distributions, and the sampling plan becomes a stratified sampling plan. Furthermore, since the strata are all of equal size and the same number of observations is drawn from each stratum, the sample is a proportionally allocated stratified random sample. The standard results from sampling theory carry over. The stratified sampling estimate of μ happens to be the sample mean of the measured units, $\bar{Y}_{rss}$. Provided a second moment exists, $\text{var}(\bar{Y}_{rss}) \leq \text{var}(\bar{Y}_{srs})$, with strict inequality if the means of the strata differ. The reduction in variance can be substantial if the quality of ranking is good.

McIntyre's early work has given rise to a growing literature on the benefits of ranked set sampling as compared to simple random sampling. Some of this research concentrates on nonparametric estimation of population characteristics, such as the mean Takahasi and Wakimoto (1968), Dell and Clutter (1972), Stokes (1977); the variance Stokes (1980), Sinha *et al.* (1996), Yu *et al.* (1999), MacEachern *et al.* (2002); the location shift between two populations Bohn and Wolfe (1992), Ozturk and Wolfe (2000*a*, 2000*b*); and the entire distribution function Stokes and Sager (1988), Kvam and Samaniego (1994), Ozturk (2002). Kaur *et al.* (1995) provided an annotated bibliography of the literature on ranked set sampling. Since 1995, there has been an increased research activity in the allocation of order statistics in a ranked set sample. See Barnett (1999), Ozturk and Wolfe (2000*c*), Chen (2001), Kaur *et al.* (1997) to name a few.

Several authors have examined the impact of imperfect judgement on estimation and testing, in a ranked set sample setting Dell and Clutter (1972), Stokes (1977), Bohn and Wolfe (1994), Barnett and Moore (1997), Nahhas *et al.* (2002), Stark (2001). This research uniformly indicates that an increase in judgement ranking error decreases the efficiency of the statistical procedure based on a ranked set sample with respect to a corresponding procedure based on a simple random sample. On the other hand, the efficiency of the ranked set sample procedure is never worse than the simple random sample procedure, even when rankings are completely random. In this worst case of random rankings, both sampling methods yield procedures that have the same efficiency, indicating that nothing is lost by using a ranked-set sampling procedure.

There is one persistent criticism of ranked set sampling in comparison to simple

random sampling. Namely, that a ranked set sample of size n relies on a greater number of experimental units than does a corresponding simple random sample. With a set size of k, a balanced ranked set sample will make the cheap measurement on nk units and the gold-standard measurement on n units; the simple random sample makes only the gold-standard measurement on n units. Nahhas (1999), Nahhas *et al.* (2002) and Stark (2001) have studied the impact of cost, and have found that it is beneficial to collect a ranked set sample under a broad range of costs. Nevertheless, there are some settings where one wishes to use every available unit; for example in many clinical trials, it is difficult to recruit enough patients for the study. In this circumstance, the limiting factor is the number of experimental units rather than cost.

In seeking to reduce the number of units used for cheap measurement, several groups have investigated making gold-standard measurements on more than one unit per set Ozturk and Wolfe (2000*d*), Bhoj (1997). For the one-sample problem described above, if one samples n/k sets and measures all units in each set, the sampling plan becomes identical to a simple random sample of size n. Thus all benefits of the ranked set sample are lost. The mechanism by which the benefits are lost is the positive correlation between the gold-standard measurements of units in the same set. We propose to exploit this positive correlation in multiple sample problems through pre-experimental judgements about the units. These judgements are then used to assign members of the same set to different treatments. To estimate a difference in treatment means with a set size of two, the positive correlation of observations within a set, and hence of the two treatment's sample means, becomes a negative correlation for the difference of the two treatment's sample means, thus reducing the variance of the estimated contrast.

The benefits that we see for contrasts extend to sign procedures and to rank procedures. In this paper, we examine the implications of a full-use ranked set sample on such "nonparametric" procedures. Section 2 outlines the model and randomization plan for the m-sample ($m \geq 2$) problems, Section 3 develops the sign test, Section 4 describes the setting of multiple treatments versus a control. Section 5 discusses practical implementation of the procedure when perfect rankings are suspect, and Section 6 provides simulations that document the benefits of full-use of the ranked set sample. Finally, Section 7 provides concluding remarks.

2. Model and randomization

In this section, we describe a data collection procedure that improves the statistical inference in an anova model with m treatments $m \geq 2$. Let $F(x-\theta_i)$, $i = 1, \ldots, m$, be the i-th treatment population having the unique median θ_i. Throughout the remainder of this paper, we reserve the subscripts C and 1 for the control group. Subscripts $2, \ldots, m$ denote the treatment groups. We wish to test

$$H_0 : \theta_1 = \theta_i; \qquad i = 2, \ldots, m$$

against the alternative hypothesis

$$H_A : \theta_1 \neq \theta_i \text{ at least for one } i,\ i \geq 2$$

or

$$H_{A'} : \theta_1 < \theta_i \text{ at least for one } i,\ i \geq 2.$$

Note that the alternative hypothesis H_A corresponds to a two-sided test whereas $H_{A'}$ corresponds to a one-sided test. Procedures for hypothesis H_A are well-known for a simple random sample Hettmansperger (1984), Chakraborti and Hettmansperger (1996) and has been investigated for a ranked set sample in Ozturk (1999*a*) and Ozturk *et al.* (2004).

To test this hypothesis, we use a variant on ranked set sampling, that we call full-use ranked set sampling, to assign the units to the two treatments. For simplicity, we first assume that we have two sample problems ($m = 2$). In this case, for a full-use ranked set sample, $2n$ pairs of experimental units are obtained for use in the study. The two units in each pair are ranked as the "smallest" (rank 1) and "largest" (rank 2), based on some perhaps subjective general quality measure of the units. Since the units have not yet been treated, this is an assessment of the unit itself rather than the response of a unit, as is traditional in ranked set sampling. A simple random sample of n of the $2n$ sets is chosen. In these sets, the unit ranked 1 is assigned to the control, and the unit ranked 2 is assigned to the treatment. For the remaining n sets, the unit ranked 1 is assigned to the treatment and the unit ranked 2 is assigned to the control. Let $Y_{C[j]k}$ be the response variable in the control group, from the j-th ranked unit in the k-th set, $j = 1, 2$; $k = 1, \ldots, n$. Similarly, $Y_{T[j]k}$ denote the response variable in the treatment group, from the j-th ranked unit in the k-th set, $j = 1, 2$; $k = 1, \ldots, n$. The design of the experiment is illustrated in Table 1.

The design just described is a balanced design, where n units of each rank are assigned to the treatment. A more general class of designs allows unbalanced designs, including those where the number of treatment units with each rank is random. For example, a complete randomization may be used, where $2n$ independent Bernoulli(0.5) variates determine whether a set is of the first type (i.e., assigns the lower ranked unit to the control and the higher ranked unit to the treatment) or of the second type (i.e., assigns the lower ranked unit to the treatment and the higher ranked unit to the control). Conditioning on the observed number of sets of the two types, inference proceeds as if these numbers were fixed by design. When the number of each type of set is fixed by design, we refer to the full-use ranked set sample as an order restricted randomized (*orr*) design. The work to follow focuses on balanced, order restricted, randomized designs.

When $m > 2$, for treatment versus control comparisons, all $m - 1$ contrasts will be of the form $\theta_i - \theta_1$, $i = 2, \ldots, m$. To focus on these contrasts, we select a set size of 2 and choose a design that is balanced across the treatments. Each set will contain the control

Table 1. Sampling design for a two sample test for the contrast parameter $\Delta = \theta_C - \theta_T$.

Set	Rank	Treatment	Response	Rank	Treatment	Response
1	1	Control	$Y_{C[1]1}$	2	Treatment	$Y_{T[2]1}$
2	1	Control	$Y_{C[1]2}$	2	Treatment	$Y_{T[2]2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	Control	$Y_{C[1]n}$	2	Treatment	$Y_{T[2]n}$
1	1	Treatment	$Y_{T[1]1}$	2	Control	$Y_{C[2]1}$
2	1	Treatment	$Y_{T[1]2}$	2	Control	$Y_{C[2]2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	Treatment	$Y_{T[1]n}$	2	Control	$Y_{C[2]n}$

and a single treatment. In this case, the basic design in Table 1 (T being replaced with i), is replicated for each treatment i, $i = 2, \ldots, m-1$.

The following remarks provide important features of these designs:

1- For each treatment condition, the design produces a balanced ranked set sample since each treatment (and the control) is repeated an equal number of times within each rank.

2- The observations obtained from the same set are not independent. They are, in general, positively correlated.

3- All observations coming from different sets are independent.

4- When $m > 2$, the control group will have $N_1 = 2\sum_{i=2}^{m} n_i$ observations whereas the i-th treatment group, $i = 2, \ldots, m$, will have $N_i = 2n_i$ observations.

The analyses that we propose in this paper are motivated by the additive model (see, for example, Ramsey and Schaefer (2002)). The additive model states that the treatments and control are members of a location family with cdf F when the median is 0. An observation from a member of the family with median θ is $Y = \theta + \epsilon$, where $\epsilon \sim F$. The value ϵ is attached to the experimental unit so that, if assigned to a treatment with median θ_i, the measured value for the unit becomes $Y_i = \theta_i + \epsilon$.

Throughout the paper we use $f_{[j]}(t-\theta_i)$ and $F_{[j]}(t-\theta_i)$ to denote the pdf and cdf of the j-th judgement order statistics of size 2 from the i-th treatment population, $i = 1, \ldots, m$. A square brackets subscript indicates possible ranking errors. A round brackets subscript indicates perfect rankings. Thus, ranking the two members of a set induces a distribution on the two ranking classes, so that $Y_{[j]} = \theta + \epsilon_{[j]}$ for $j = 1, 2$. If rankings are perfect, the vector $(\epsilon_{[1]}, \epsilon_{[2]})$ is distributed as $(\epsilon_{(1)}, \epsilon_{(2)})$; if rankings are imperfect, the order statistics of the vector $(\epsilon_{[1]}, \epsilon_{[2]})$ are distributed as $(\epsilon_{(1)}, \epsilon_{(2)})$. For the i-th treatment group responses $Y_{i[j]k}$; $j = 1, 2; k = 1, \ldots, n$ are independent but not identically distributed; however $Y_{i[j]k}$, $k = 1, \ldots, n$, are identically distributed with pdf $f_{[j]}(t-\theta_i)$ for fixed j.

3. One sample sign statistic

The control and treatment samples, considered individually, are balanced ranked set samples. The one-sample sign test has been developed for this sampling plan in Hettmansperger (1995), Koti and Babu (1996), and Ozturk (1999a). Results and notation follow.

Let

$$\bar{S}_i(\theta) = \frac{1}{2n} S_i^+(\theta) - 1/2 = \frac{1}{2n} \sum_{k=1}^{n} \sum_{j=1}^{2} \{I(Y_{i[j]k} - \theta) - 1/2\}, \qquad i = 1, 2,$$

where $I(a) = 0, 1$ as $a \leq, > 0$. Note that $S_1^+(\theta)$ and $S_2^+(\theta)$ are sign statistics for the control and treatment groups. They are monotone decreasing functions of θ. Therefore, the one-sample sign test rejects $\theta_1 = 0$ ($\theta_2 = 0$) against $\theta_1 \neq 0$ ($\theta_2 \neq 0$) if $S_1^+(0)$ ($S_2^+(0)$) is too small or too large. As with a simple random sample, we can use the monotonicity of $S_i^+(0)$ to construct a $100 \times (1-\alpha_i)\%$ confidence interval for θ_i. Let $Y^*_{1(1)} < Y^*_{1(2)} < \cdots < Y^*_{1(2n)}$ be the ordered values of $Y_{1[j]k}$; $k = 1, \ldots, n$; $j = 1, 2$. Similarly, let $Y^*_{2(1)} < Y^*_{2(2)} < \cdots < Y^*_{2(2n)}$ denote the ordered values of $Y_{2[j]k}$; $k = 1, \ldots, n$; $j = 1, 2$. Then

$$I_i = [Y^*_{i(d_i+1)}, Y^*_{i(2n-d_i)}] = [L_i, U_i],$$

is a $100 \times (1 - \alpha_i)\%$ confidence interval for θ_i, $i = 1, 2$ where $P(S_i^+(0) \leq d_i) \approx \alpha_i/2$. Because of the discrete nature of $S_i^+(0)$ exact $1-\alpha_i$ confidence level may not be achieved.

The choices of the depths d_i, $i = 1, 2$, for a specified confidence interval require one to know the exact distributions of $S_i^+(0)$. The exact distribution of this statistic is given in Koti and Babu (1996) in the context of a balanced ranked set sample and in Ozturk (1999*b*) in the context of an unbalanced ranked set sample. We state the following theorem without proof.

THEOREM 3.1. *Assume that the units are ranked perfectly and that the ϵ_j are drawn from a continuous distribution, F. Further assume that the data have been collected with the balanced design described in Section 2, and that the additive model holds. Then the exact distribution of $S_i^+(0)$, under the assumption that $\theta_1 = 0$ and $\theta_2 = 0$, is given by*

$$P(S_i^+(0) = t) = \sum_{\boldsymbol{J}_t} \prod_{r=1}^{2} \binom{n}{j_r} [1 - I_{1/2}(r, 3-r)]^{j_r} [I_{1/2}(r, 3-r)]^{n-j_r} \tag{3.1}$$

for $t = 0, 1, \ldots, 2n$ where

$$I_x(a, b) = \frac{1}{B(a, b)} \int_0^x u^{a-1}(1-u)^{b-1} du,$$
$$\boldsymbol{J}_t = \{(j_1, j_2) : j_1 + j_2 = t, 0 \leq j_r \leq n, r = 1, 2\},$$

and $B(a, b)$ is the beta function.

This theorem extends to imperfect rankings. To do so, replace $I_{1/2}(r, 3-r)$ with $F_{[r]}(0)$ in (3.1). By using the extended version of the theorem, it is trivial to construct distribution free confidence intervals for the population medians. The following results follow from Hettmansperger (1995) and Ozturk(1999*a*). Without loss of generality, we take $\theta_i = 0$, $i = 1, 2$.

$$E\bar{S}_i(0) = 0, \qquad \text{var}(\bar{S}_i(0)) = \frac{1}{2n} \sum_{j=1}^{2} F_{[j]}(0)\{1 - F_{[j]}(0)\}/2 = \frac{\bar{\delta}^2}{2n}, \qquad i = 1, 2.$$

The quantity $\sqrt{2n}\bar{S}_i(0)/\bar{\delta}$ is asymptotically distributed as the standard normal. Relying on asymptotic normality, we obtain

$$\begin{aligned} P(S_i^+(0) \leq d_i + 0.5) &= P\left(\sqrt{2n}\bar{S}_i(0)/\bar{\delta} \leq \frac{\sqrt{2n}\{(d_i + 0.5)/(2n) - 1/2\}}{\bar{\delta}}\right) \\ &\approx P\left(Z \leq \frac{\sqrt{2n}\{(d_i + 0.5)/(2n) - 1/2\}}{\bar{\delta}}\right). \end{aligned}$$

The continuity corrected depth d_i becomes

$$d_i = 2n/2 - \sqrt{2n}\bar{\delta} z_{1-\alpha_i/2} - 0.5, \qquad i = 1, 2,$$

where z_b is the b-th quantile of the standard normal distribution.

THEOREM 3.2. *Let $\theta_1 = \theta_2 = 0$. Suppose $F(t)$ is differentiable, then for any $B > 0$*

$$\sup_{|b|\leq B} \left| \sqrt{2n}\bar{S}_i\left(\frac{b}{\sqrt{2n}}\right) - \sqrt{2n}\bar{S}_i(0) + f(0)b \right| \xrightarrow{P_0} 0, \qquad i = 1, 2,$$

where P_0 denotes convergence in probability when $\theta_i = 0$, $i = 1, 2$.

The proof of this theorem is given in Ozturk (1999*a*) for a balanced ranked set sample with perfect ranking. The proof in this case is essentially the same, so it is omitted here. In order to use Theorem 3.2, we need to show that $\sqrt{2n}L_i$ and $\sqrt{2n}U_i$ are bounded in probability.

LEMMA 3.1. *Under the assumptions of Theorem 3.2, and also taking $f(0) > 0$, for $\epsilon > 0$ there exists a $B < \infty$, such that*

$$\text{(i)} \quad \liminf_{n\to\infty} P(\sqrt{2n}|L_i| \leq B) \geq 1 - \epsilon$$

and

$$\text{(ii)} \quad \liminf_{n\to\infty} P(\sqrt{2n}|U_i| \leq B) \geq 1 - \epsilon.$$

From the asymptotic normality of the sign statistic, we can establish that $\lim_{n\to\infty} \sqrt{2n}\bar{S}_i(L_i)/\bar{\delta} = z_{1-\alpha_i/2}$ and that $\lim_{n\to\infty} \sqrt{2n}\bar{S}_i(U_i)/\bar{\delta} = z_{\alpha_i/2}$, for $i = 1, 2$. In the previous theorem, by replacing $b/\sqrt{2n}$ with L_i and U_i, and solving for these quantities we obtain

$$L_i = -\frac{z_{1-\alpha_i/2}\bar{\delta}}{\sqrt{2n}f(0)} + \frac{\bar{S}_i(0)}{f(0)} + o_p(1/\sqrt{2n}), \tag{3.2}$$

and

$$U_i = \frac{z_{1-\alpha_i/2}\bar{\delta}}{\sqrt{2n}f(0)} + \frac{\bar{S}_i(0)}{f(0)} + o_p(1/\sqrt{2n}). \tag{3.3}$$

The expressions in equations (3.2) and (3.3) are used to establish the asymptotic properties of the tests and intervals presented in subsequent sections. When $\theta_1(\theta_2) \neq 0$, the interval is shifted up or down by $\theta_1(\theta_2)$.

4. Control versus treatment multiple comparisons

In this section, our interest is to test whether any one of the $m - 1$ treatment group means differs from the control group mean. Let

$$Y_{i[j]k} = \theta_i + \epsilon_{i[j]k}, \qquad i = 1, \ldots, m; \ j = 1, 2; \ k = 1, \ldots, n_i,$$

be the response measurement from the i-th treatment on the j-th ranked unit and k-th replication.

As in the one sample problem, let $Y^*_{1(1)} < Y^*_{1(2)} \cdots < Y^*_{1(N_1)}$ be the ordered values of the N_1 observations from the control group and let $Y^*_{i(1)} < Y^*_{i(2)} \cdots < Y^*_{i(N_i)}$ be the ordered values of the N_i observations from the i-th treatment group. Then $100(1-\alpha_i)\%$ confidence intervals for θ_i, $i = 1, \ldots, m$ can be constructed as

$$I_i = [Y^*_{i(d_i+1)}, Y^*_{i(N_i-d_i)}] = [L_i, U_i],$$

where $P(S_i(0) \leq d_i) = \alpha_i/2$, and

$$S_1(a) = \frac{1}{N_1} \sum_{j=1}^{2} \sum_{k=1}^{n_1} I(Y_{1[j]k} - a),$$

$$S_i(a) = \frac{1}{N_i} \sum_{j=1}^{2} \sum_{k=1}^{n_i} I(Y_{i[j]k} - a), \qquad i = 2, \ldots, m.$$

Rejection rule I. We reject the null hypothesis H_0 in favor of H_A if, for some $i \geq 2$, I_1 and I_i are disjoint.

Rejection rule II. We reject the null hypothesis H_0 in favor of $H_{A'}$ if, for some $i \geq 2$, I_1 and I_i are disjoint and I_1 is located to the left of I_i.

Rejection rule I provides a test for the two-sided alternative H_A whereas Rejection rule II provides a test for the one-sided alternative $H_{A'}$. Note that both rejection rules make $m-1$ comparisons. To preserve the level of the test, we must match the familywise error rate to the specified level. The familywise error rates for Rejection rule I and Rejection rule II can be written as follows

$$\alpha_E = P_{H_0}(I_1 \text{ and } I_i \text{ are disjoint for at least one } i \geq 2)$$

and

$$\alpha_{E'} = P_{H_0}(I_1 \text{ and } I_i \text{ are disjoint and } I_1 \text{ is on the left of } I_i \text{ for at least one } i \geq 2).$$

Let $\lambda_i = N_i/N$, where $N = N_1 + \sum_{i=2}^{m} N_i$. Then, by using the linear representations in Section 3, we obtain

$$\begin{aligned} \alpha_E &= P_{H_0}(L_i - U_1 > 0 \text{ or } U_i - L_1 < 0 \text{ for at least one } i, i \geq 2) \\ &= P_{H_0}(|T_i| > D_i \text{ for at least one } i, i \geq 2) \\ &\leq P_{H_0}\left(\max_{2 \leq i \leq m} |T_i| > \min_{2 \leq i \leq m} D_i\right) \end{aligned} \tag{4.1}$$

and

$$\alpha_{E'} \leq P_{H_0}\left(\max_{2 \leq i \leq m} T_i > \min_{2 \leq i \leq m} D_i\right) \tag{4.2}$$

where

$$T_i = \frac{\sqrt{N}\{\bar{S}_i(0) - \bar{S}_1(0)\}}{\sqrt{\dfrac{\bar{\delta}^2}{\lambda_1} + \dfrac{\bar{\delta}^2}{\lambda_i} - \dfrac{2\delta_{[1,2]}}{\lambda_1}}}, \qquad D_i = \frac{z_{1-\alpha_i/2}\bar{\delta}\lambda_i^{-1/2} + z_{1-\alpha_1/2}\bar{\delta}\lambda_1^{-1/2}}{\sqrt{\dfrac{\bar{\delta}^2}{\lambda_1} + \dfrac{\bar{\delta}^2}{\lambda_i} - \dfrac{2\delta_{[1,2]}}{\lambda_1}}}, \qquad i = 2, \ldots, m.$$

and

$$\delta_{[1,2]} = F_{[1,2]}(0,0) - F_{[1]}(0)F_{[2]}(0).$$

In both expressions above, equality holds if all of the D_i are equal and inequality holds if some of the D_i differ. It is clear that the familywise error rates are determined by the distribution of $\boldsymbol{T} = (T_2, \ldots, T_m)$.

THEOREM 4.1. *Under H_0, as $N \to \infty$, $\boldsymbol{T}$ converges to an $(m-1)$ dimensional normal random variable with mean zero, variances* $\operatorname{var}(T_i) = 1$, $i = 2, \ldots, m$ *and covariances*

$$\operatorname{cov}(T_i, T_j) = r_{i,j} = \frac{\bar{\delta}^2/\lambda_1 - \delta_{[1,2]}/\lambda_1}{\sqrt{\frac{\bar{\delta}^2}{\lambda_1} + \frac{\bar{\delta}^2}{\lambda_i} - \frac{2\delta_{[1,2]}}{\lambda_1}}\sqrt{\frac{\bar{\delta}^2}{\lambda_1} + \frac{\bar{\delta}^2}{\lambda_j} - \frac{2\delta_{[1,2]}}{\lambda_1}}}, \quad i \neq j \quad \text{and} \quad i, j \geq 2.$$

Remarks:

1. If $\lambda_i = \lambda_j$, for all $i, j \geq 2$, then

$$r_{i,j} = r = \frac{\bar{\delta}^2 - \delta_{[1,2]}}{m\bar{\delta}^2 - 2\delta_{[1,2]}}.$$

2. If $\delta_{[1,2]} = 0$, then the asymptotic distribution of $\boldsymbol{T}$ in *orr* is the same as the distribution of $\boldsymbol{T}$ in a balanced ranked set sample and a simple random sample. In this case, $r^*_{i,j}$, the correlation coefficient between T_i and T_j, does not depend on $\bar{\delta}$. It is

$$r^*_{i,j} = \frac{{\lambda_1}^{-1}}{\sqrt{{\lambda_1}^{-1} + {\lambda_i}^{-1}}\sqrt{{\lambda_1}^{-1} + {\lambda_j}^{-1}}}.$$

Let γ_u be the u-th quantile of the $\max(\boldsymbol{T}_A)$, where $\boldsymbol{T}_A = (|T_2|, \ldots, |T_m|)$. Then, for a given value of the familywise error rate, α_E, we need to determine the $(1-\alpha_E)$-th quantile of the $\max(\boldsymbol{T}_A)$. Although the asymptotic distribution of $\boldsymbol{T}$, under perfect and random judgement, does not depend on the underlying distribution $F(\cdot)$, it is not readily available for tabulation. The following lemma provides a computational formula for the u-th quantile of the $\max(\boldsymbol{T}_A)$.

LEMMA 4.1. *Assume that $\lambda_i = \lambda$, for all $i \geq 2$. Let $T^* = \max(\boldsymbol{T}_A)$. Then the u-th quantile of the asymptotic distribution of T^*, γ_u, can be obtained as a solution of the following equation*

$$u = \int \left\{ \Phi^{m-1}\left(\frac{\gamma_u - x}{\sqrt{1-r}}\right) - \Phi^{m-1}\left(-\frac{\gamma_u + x}{\sqrt{1-r}}\right) \right\} \phi(x/\sqrt{r})/\sqrt{r}dx.$$

In a similar fashion, we also evaluate the asymptotic quantiles of $T^+ = \max(\boldsymbol{T})$ for a one-sided test.

LEMMA 4.2. *Assume that $\lambda_i = \lambda$, for all $i \geq 2$. Let $T^+ = \max(\boldsymbol{T})$. Then the u-th quantile of the asymptotic distribution of T^+, γ^+_u, can be obtained as a solution of the following equation*

$$u = \int \Phi^{m-1}\left(\frac{\gamma^+_u - x}{\sqrt{1-r}}\right) \phi(x/\sqrt{r})/\sqrt{r}dx.$$

The tabulated quantiles of T^* and T^+ depend on the correlation coefficient r. The correlation coefficient depends on the judgement ranking process. We do know, however,

Table 2. The $(1-\alpha_E)$-th quantiles of the asymptotic distributions of T^* and T^+ for perfect judgment ranking and different values of m.

	Two-sided		One-sided			Two-sided		One-sided	
m	$\gamma_{0.95}$	$\gamma_{0.99}$	$\gamma^+_{0.95}$	$\gamma^+_{0.99}$	m	$\gamma_{0.95}$	$\gamma_{0.99}$	$\gamma^+_{0.95}$	$\gamma^+_{0.99}$
3	2.230	2.804	1.940	2.570	5	2.487	3.021	2.223	2.803
4	2.382	2.932	2.108	2.708	6	2.566	3.088	2.310	2.874
4	2.379	2.931	2.103	2.706	6	2.563	3.087	2.304	2.872

that when the additive model holds, the correlation varies between $1/m$ (random judgement) and $2/(3m-2)$ (perfect judgement). Numerical computations have shown that quantiles of T^* and T^+ did not change much within the range from random judgement to perfect judgement. The maximum discrepancy $(\gamma(perfect) - \gamma(random))$ is less than or equal to 0.002, which can be ignored for practical purposes. Therefore, in Table 2, we tabulate the 95th and 99th percentiles of T^* and T^+ for perfect judgement ranking.

For a given Type I error rate, α_E, an upper bound on the familywise error rate can now be determined from the individual confidence coefficients for the control and treatment confidence intervals. For a fixed α_E, from equations (4.1) and (4.2) we have that

$$\gamma_{\alpha_E} = \min_{2\le i\le m} D_i = \min_{2\le i\le m}\left\{\frac{z_{1-\alpha_i/2}\bar{\delta}\lambda_i^{-1/2} + z_{1-\alpha_1/2}\bar{\delta}\lambda_1^{-1/2}}{\sqrt{\dfrac{\bar{\delta}^2}{\lambda_1}+\dfrac{\bar{\delta}^2}{\lambda_i}-\dfrac{2\delta_{[1,2]}}{\lambda_1}}}\right\} \tag{4.3}$$

and

$$\gamma^+_{\alpha_E} = \min_{2\le i\le m} D_i = \min_{2\le i\le m}\left\{\frac{z_{1-\alpha_i/2}\bar{\delta}\lambda_i^{-1/2} + z_{1-\alpha_1/2}\bar{\delta}\lambda_1^{-1/2}}{\sqrt{\dfrac{\bar{\delta}^2}{\lambda_1}+\dfrac{\bar{\delta}^2}{\lambda_i}-\dfrac{2\delta_{[1,2]}}{\lambda_1}}}\right\}. \tag{4.4}$$

In our simulations, we have used equal confidence coefficients for the control and each of the treatments. Let $Z = z_{1-\alpha_i/2} = z_{1-\alpha_1/2}$, for $i = 2,\ldots,m$. If the sample sizes are all equal in the treatment groups, we have that $\lambda_1 = 1/2$ and $\lambda_i = 1/(2(m-1))$, $i = 2,\ldots,m$. Then, putting these in equations (4.3) and (4.4) and solving for Z, we obtain

$$Z = \frac{\gamma_{\alpha_E}\sqrt{m\bar{\delta}^2 - 2\delta_{[1,2]}}}{\bar{\delta}\{\sqrt{m-1}+1\}} \tag{4.5}$$

and

$$Z^+ = \frac{\gamma^+_{\alpha_E}\sqrt{m\bar{\delta}^2 - 2\delta_{[1,2]}}}{\bar{\delta}\{\sqrt{m-1}+1\}}.$$

The individual confidence coefficient can then be determined by the relation $1-\alpha_i = 1-2\Phi(-Z)$, $i = 1,\ldots,m$, for the two-sided test and $1-\alpha_i = 1-\Phi(-Z^+)$, $i = 1,\ldots,m$, for the one-sided test. Table 3 provides the individual confidence coefficients for 5% and 1% two-sided tests, under perfect and random ranking, when $m = 2,3,4,5,6$.

When $m = 2$, an alternative construction of the proposed test can be considered directly from the confidence interval of the difference parameter Δ. We note that $[L_1 - U_2, U_1 - L_2]$ is a $100(1 - \alpha_I)\%$ confidence interval for Δ. Thus, we reject the null hypothesis H_0 if zero is not contained in the confidence interval for the difference parameter which is equivalent to the decision rule that rejects the null hypothesis if the individual median confidence intervals do not overlap. Then, the asymptotic efficiency of the proposed test is measured in terms of asymptotic length of the confidence interval for Δ. Readers are referred to Randles and Wolfe (1991) for details of the efficiency of tests based on confidence intervals.

THEOREM 4.2. *Assume $m = 2$, and let $N = 4n$ and define the width of the confidence interval based on N observations to be $w_N = (U_1 - L_2 - L_1 + U_2)$. Under the conditions of Theorem* 3.2, *as $n \to \infty$, $N^{1/2}w_N$ converges in probability to w^*, where*

$$w^* = \frac{4(\bar{\delta}^2 - \delta_{[1,2]})^{1/2}\gamma_{\alpha_E}}{f(0)}.$$

Theorem 4.2 and equation (4.3) indicate that although there are infinitely many $z_{1-\alpha_1/2}$ and $z_{1-\alpha_2/2}$ for a given γ_{α_E}, the asymptotic length of the confidence interval does not depend on them. All choices produce the same asymptotic length. This provides a flexibility in practice. A researcher can choose any $z_{1-\alpha_1/2} > 0$ and $z_{1-\alpha_2/2} > 0$ as long as they produce the desired Type I error rate.

In order to compare the proposed procedure with its competitors in the literature, we use the asymptotic Pitman efficacy of the test. The quantity $(\sqrt{N}w/(2\gamma_{\alpha_E}))^2$ converges to the reciprocal of the Pitman efficacy of the confidence-interval-based test. Thus, the Pitman efficacy of the order restricted randomized (orr) two-sample median test is

$$\text{eff}(orr) = \frac{f^2(0)}{4\{\bar{\delta}^2 - \delta_{[1,2]}\}}.$$

The Pitman efficacies of the ranked set sample and simple random sample two-sample median tests are given in Ozturk (1999a)

$$\text{eff}(rss) = \frac{f^2(0)}{4\bar{\delta}^2}, \qquad \text{eff}(srs) = f^2(0).$$

We now obtain the Pitman asymptotic relative efficiency of orr with respect to rss and srs

$$ARE(orr, rss) = \text{eff}(orr)/\text{eff}(rss) = \frac{1}{1 - \dfrac{\delta_{[1,2]}}{\bar{\delta}^2}}$$

$$ARE(orr, srs) = \frac{1}{4(\bar{\delta}^2 - \delta_{[1,2]})}.$$

Note that, under the additive model, $\delta_{[1,2]} \geq 0$. It is then immediate to observe that $ARE(orr, rss) \geq 1$.

Remark 1. Under perfect ranking, $ARE(orr, rss) = 1.5$ and $ARE(orr, srs) = 2$. This compares to $ARE(rss, srs) = 4/3$.

5. Implementation of the procedure

For the implementation of the proposed procedure, we need to determine individual confidence coefficients which depend on the quantiles of T^*, T^+ and the parameters $\bar{\delta}$ and $\delta_{[1,2]}$. The quantiles change only slightly with the quality of judgement ranking, and so we use the asymptotic quantiles for perfect judgement ranking in Table 2 in all cases. On the other hand, the parameters $\bar{\delta}$, $\delta_{[1,2]}$ and the individual confidence coefficients can change dramatically with the quality of ranking, so must be estimated from the data. Let $\hat{\theta}_i$, $i = 1, \ldots, m$, be the sample medians of the control and treatment groups. In order to provide a pooled estimator, we center the observations by subtracting the appropriate sample median from each observation; $Z^*_{1[j]k} = Y_{1[j]k} - \hat{\theta}_1$; $k = 1, \ldots, n_1$; $j = 1, 2$ and $Z^*_{i[j]k} = Y_{i[j]k} - \hat{\theta}_i$; $k = 1, \ldots, n_i$; $j = 1, 2$; $i = 2, \ldots, m$. Let $Z_{[1]k}$, $k = 1, \ldots, N/2$, and $Z_{[2]k}$, $k = 1, \ldots, N/2$, be the pooled centered observations of the first and second judgement order statistics over all treatment and control groups. For example, in this notation, $Z_{[1]k}$, $k = 1, \ldots, N/2$, represent the collection of all $Z^*_{i[1]k}$, $k = 1, \ldots, n_i$; $i = 1, \ldots, m$. Now we define

$$\hat{F}_{[j]}(t) = \frac{2}{N} \sum_{k=1}^{N/2} I^*(Z_{[j]k} - t), \qquad j = 1, 2.$$

$$\hat{F}_{[1,2]}(t, s) = \frac{2}{N} \sum_{k=1}^{N/2} I^*(Z_{[1]k} - t, Z_{[2]k} - s)$$

where $I^*(a) = 1, 0$ as $a \leq, > 0$, $I^*(a, b) = 1$ if $a \leq 0$, $b \leq 0$ and zero otherwise. We estimate $\bar{\delta}^2$ with

$$\hat{\delta}^2 = \frac{1}{2} \sum_{j=1}^{2} \hat{F}_{[j]}(0)(1 - \hat{F}_{[j]}(0))$$

and $\delta_{[1,2]}$ with

$$\hat{\delta}_{[1,2]} = \hat{F}_{[1,2]}(0, 0) - \hat{F}_{[1]}(0)\hat{F}_{[2]}(0).$$

In practice, we first estimate $\bar{\delta}^2$ and $\delta_{[1,2]}$ and then compute the value of Z, $Z = z_{\alpha_i/2}$; $i = 1, \ldots, m$, which corresponds to the individual confidence coefficients for a given value of Type I error rate. We construct the median confidence intervals either by using the asymptotic results or by appealing to the exact distribution of the sign statistic in Theorem 3.1 and subsequent paragraph. We, of course, replace $I_{1/2}(r, 3-r)$ with $\hat{F}_{[r]}(0)$, a consistent estimate of $F_{[r]}(0)$, $r = 1, 2$. Then, we proceed with the decision rule that corresponds to the hypothesis of interest. The asymptotic results for the one-sample, two-sample and m-sample problems are unaffected by consistent estimation of $\bar{\delta}$ and $\delta_{[1,2]}$. Consistency of $\hat{\delta}$ and $\hat{\delta}_{[1,2]}$ for $\bar{\delta}$ and $\delta_{[1,2]}$ follows from the consistency of $\hat{F}_{[j]}(0)$ and $\hat{F}_{[1,2]}(0, 0)$ for $F_{[j]}(0)$ and $F_{[1,2]}(0, 0)$, respectively.

6. Empirical power and simulation study

This section looks at the empirical power of the proposed testing procedure under different sampling schemes and under varying quality of judgement ranking. The simulation study considers simple random samples, ranked set samples and order restricted randomized samples.

The simulation relies on a model for the judgement ranking process. Specifically, we follow Dell and Clutter's (1972) bivariate normal model of X and ϵ with $\text{corr}(X, \epsilon) = \rho$. Both X and ϵ are marginally standard normal. In order to generate judgement ranked order statistics, we first generate a pair of bivariate normal vectors, (X_1, ϵ_1), (X_2, ϵ_2). The vectors are sorted on their first (X) coordinate, yielding $\epsilon_{[1]}$ and $\epsilon_{[2]}$. The $\epsilon_{[j]}$ are then turned into responses by adding the appropriate treatment median. When $\rho = \text{corr}(X, \epsilon) = 1$, ranking is perfect, and $\epsilon_{[1]} < \epsilon_{[2]}$; when $\rho = 0$, ranking is random, and $P(\epsilon_{[1]} < \epsilon_{[2]}) = 0.5$. For $0 < \rho < 1$, ranking of intermediate quality can be investigated.

In the simulation, the number of treatments are taken as $m = 2, 3$, the sample sizes in the treatment groups are taken as $n_i = 10, 15, 20, 30$, for $i = 2, \ldots, m$, and the familywise error rate is taken as 0.05 and 0.01. To save space, we only report the results for $N_i = 10, 30$. In order to evaluate the empirical power of the test, the control samples are shifted by adding the shift parameters ($shift = 0(2)0.1$). The empirical power of the test is computed for *orr*, ranked set and simple random samples for different values of $\rho = 0.50, 0.75, 1.00$. Figure 1 presents the empirical power of the test for a two-sample procedure at the 5% level, and Fig. 2 presents the empirical power of the test when $m = 3$. The solid lines represent the empirical power of order restricted randomized samples while the dashed lines represent the empirical power of the standard ranked set samples. The power curve of the simple random sample median test is marked with $\bigtriangledown$.

We can summarize the main features of the figures as follows:

1. The level of the test holds quite well for sample sizes as low as $n = 10$ when $m = 3$ and $n = 20$ (this is not reported due to a limitation on space) when $m = 2$. As

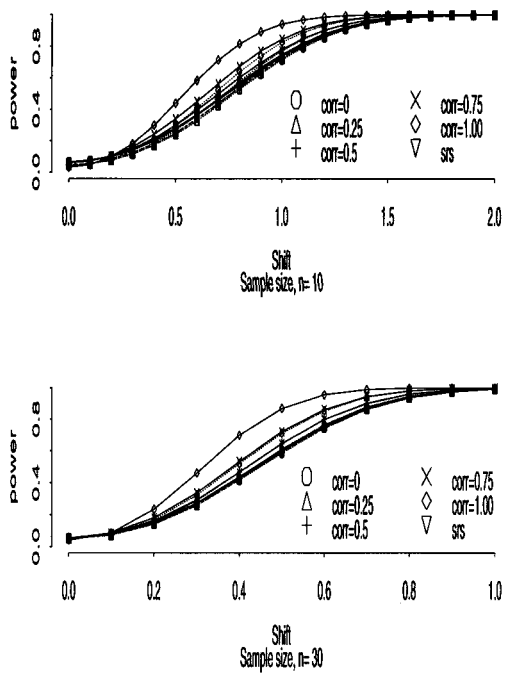

Fig. 1. Empirical powers of two-sample median tests: Solid lines represent order restricted randomizations; dashed lines represent ranked set samples; corr = $\rho = \text{corr}(X, \epsilon)$.

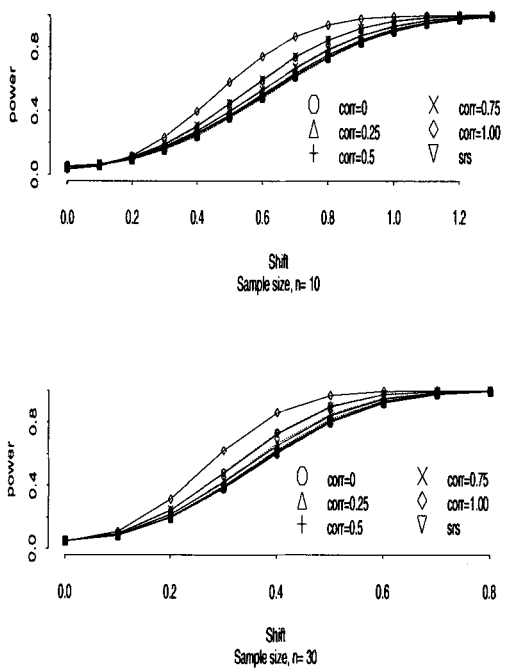

Fig. 2. Empirical powers of three-sample median tests ($m = 3$): Solid lines represent order restricted randomizations; dashed lines represent ranked set samples; corr = $\rho = \text{corr}(X, \epsilon)$.

to be expected, the actual level tends to move closer to the nominal level as the sample size increases.

2. For $\rho > 0$ there is a clear distinction between all three sampling procedures; *srs* provides the least power, *rss* has the next highest power and *orr* sample has the greatest power. The greater the ρ is, the bigger is the differential among the sampling procedures.

3. It is very clear from the figures that the *orr* samples can handle imperfect ranking better than the ranked set samples. The power curves of *orr* samples for a given value of ρ almost always lie above the power curves of ranked set samples with a $\rho^* > \rho$. For example power curves of *orr* designs with $\rho = 0.75$ stay above the power curves of ranked set sample with $\rho = 1.00$. Similar results can also be observed for other ρ values.

In the simulation study, we also estimated the coverage probabilities of the individual confidence intervals. Table 4 presents the empirical coverage probabilities in percent for intervals based on the 5% and 1% two-sided tests. As the correlation coefficient, ρ, increases, indicating better ability to rank the units, the theoretical, asymptotic individual coverage necessary to obtain the targeted familywise error rate decreases. This is a consequence of the increased correlation between the ϵ_j within a set, as ρ increases. This behavior is to be anticipated from equation (4.5). For $m = 2$, we have that

$$z = z_{1-\alpha_1/2} = z_{1-\alpha_2/2} = \frac{\{\bar{\delta}^2 - \delta_{[1,2]}\}^{1/2}}{\sqrt{2}\bar{\delta}} \gamma_{\alpha_E}.$$

Since $\delta_{[1,2]}$ is an increasing function of ρ (recall that $\rho > 0$) and γ_{α_E} does not change much with ρ, z is a decreasing function of ρ. A comparison of selected values from Tables 3 and 4 show that the individual intervals' coverage rates are very close to their

Table 3. Individual confidence coefficients (in percent) for 5% and 1% tests.

α_E	Judg.	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$
0.05	Perfect	74	84	89	92	93
0.01	Perfect	86	92	95	96	97
0.05	Random	84	89	92	94	95
0.01	Random	93	96	97	98	98

Table 4. Empirical coverage probabilities (in percent) of the individual confidence intervals; m is the number of treatment groups; n_i is the sample size in each treatment group; ρ is the correlation coefficient between the concomitant and response variable; *orr* denotes order restricted randomized sample; *rss* denotes ranked set sample.

		$m = 2$				$m = 3$			
		5% Test		1% Test		5% Test		1% Test	
n_i	ρ	*orr*	*rss*	*orr*	*rss*	*orr*	*rss*	*orr*	*rss*
10	0.00	81.26	84.06	91.49	94.26	88.76	89.28	95.49	95.71
30	0.00	82.73	83.49	92.66	93.72	89.00	89.24	95.54	95.70
10	0.25	80.98	84.08	91.31	94.29	88.64	89.31	95.40	95.76
30	0.25	82.45	83.50	92.46	93.75	88.88	89.22	95.43	95.70
10	0.50	80.14	84.15	90.72	94.40	88.21	89.33	95.17	98.87
30	0.50	81.53	83.58	91.82	93.81	88.38	89.25	95.13	95.63
10	0.75	78.39	84.26	89.58	94.45	87.44	89.50	94.60	96.02
30	0.75	79.48	83.74	90.40	93.75	87.17	89.26	94.46	95.69
10	1.00	75.43	84.36	87.35	94.45	85.48	89.87	93.03	96.07
30	1.00	74.76	83.74	86.86	93.81	84.65	89.26	92.76	95.72

targets. This indicates that asymptotic intervals for a single median, based on the sign test with estimated $\bar{\delta}$, have actual coverage very nearly equal to nominal coverage.

As a specific example, when $m = 3$ and the Type I error rate is 0.05, the empirical coverage probabilities for a single treatment median for *orr* samples are 85.48%, 84.91%, 84.91% and 84.65% respectively for perfect ranking and $n_i = 10, 15, 20$, and 30. These values are close to 84%, the theoretical value of the asymptotic coverage probability.

7. Conclusions

The literature on ranked set sampling has largely concentrated on designs that consider independent order statistics or independent judgement order statistics. A more modest segment of the literature has considered selection of multiple units from a set, but only in one-sample problems. In all cases, the research shows the superiority of ranked set samples to simple random samples with an equivalent number of fully measured units. The main drawbacks to a ranked set sample are the cost of ranking units and the fact that many of the experimental units are discarded, with no full measurement made upon them. Authors who have implemented ranked set sampling have generally found that the cost of ranking a unit is substantially less than that of making the gold-standard measurement on the unit. This research addresses the second, perhaps substantial, cost of obtaining, treating and then discarding a large number of units.

In this paper, we have described a ranked set sampling method that relies on pre-experimental judgements, and have shown that, for multiple sample problems, use of all units in a set leads to more accurate inference than does use of only some units or a single unit in a set. Thus, in terms of accuracy of inference, full-use ranked set sampling is to be preferred to ranked set sampling which is to be preferred to simple random sampling. The reduction in cost when moving from a ranked set sample to a full-use ranked set sample is an added bonus.

The theory and simulations set out in Sections 3–6 demonstrate the magnitude of improvement that is possible with the proposed sampling plan. The combination of correlated judgement order statistics and restricted randomization provides asymptotic relative efficiencies of 2.0 and 1.5 for the two-sample median test, compared to the same procedure applied to simple random sampling or ranked set sampling plans, respectively. Similar improvement is observed in the finite sample power comparisons in Figs. 1 and 2.

Implementation of the testing procedure is easy and straightforward if perfect judgement ranking holds. On the other hand, if there is some doubt on the ability of the judges to rank the units correctly, the "estimated" quality of ranking model, where $\bar{\delta}$ and $\delta_{[1,2]}$ are estimated from the data, is used to implement the procedure. The simulation study shows that the "estimated" quality of ranking model works well for moderate sample sizes. It also shows that the full-use ranked set sampling design can handle imperfect ranking better than does a standard ranked set sample design. For example, the empirical power of the order restricted design, when the correlation coefficient between the concomitant and response variable is 0.75, is better than or as good as a standard ranked set sample with perfect ranking.

The tests described here can be extended to cover some models that are more general than the additive model. Of particular interest is the location-scale family defined by $\epsilon \sim F$, with F having median 0 and scale 1. A treatment is defined by a location and scale parameter, θ and σ, respectively. The corresponding observation, if the unit is assigned to treatment i, is $Y = \theta_i + \sigma_i \epsilon$. The median based tests and confidence intervals appropriate for this model would require estimation of both treatment locations and scales which can be accomplished in a robust fashion.

In related work, the authors are investigating mean-based inference, with a focus on problems that fall in the framework of the linear model. This includes two-sample and k-sample problems as well as those generally approached through regression analysis. Preliminary results show that inference which relies on contrasts is improved with a full-use ranked set sample. This holds, whether inference is based on asymptotic theory or on a permutation distribution. The work on mean-based inference also facilitates study of robustness of the procedures to departures from the additive model.

Acknowledgements

This material is based on work supported by the National Science Foundation under Award Number DMS-0072526. The authors thank the Associate Editor and referees for a careful reading of the paper.

Appendix

PROOF OF LEMMA 3.1. The proofs of (i) and (ii) are similar. We focus on (ii). By inverting the critical region of the sign statistic, the upper limit of the $100 \times (1 - \alpha_i)\%$

confidence interval for θ_i can be written as

$$U_i = \sup\left\{t : \frac{\sqrt{2n}\bar{S}_i(t)}{\bar{\delta}} > z_{\alpha_i/2}\right\}.$$

Fix n and consider an arbitrary fixed $B > 0$. Since $\bar{S}_i(a)$ is a non-increasing function of a, when $c \leq z_{\alpha_i/2}$, the event

$$\frac{\sqrt{2n}\bar{S}_i(B/\sqrt{2n})}{\bar{\delta}} < c$$

implies that $U_i \leq B/\sqrt{2n}$. Hence,

$$P\left(\frac{\sqrt{2n}\bar{S}_i(B/\sqrt{2n})}{\bar{\delta}} < z_{\alpha_i/2}\right) \leq P(U_i \leq B/\sqrt{2n}).$$

Letting n vary and noting that $\sqrt{2n}\bar{S}_i(0)$ tends to a continuous random variable, from Theorem 3.2, we have that

$$P\left(\frac{\sqrt{2n}\bar{S}_i(B/\sqrt{2n})}{\bar{\delta}} < z_{\alpha_i/2}\right) - P\left(\frac{\sqrt{2n}\bar{S}_i(0)}{\bar{\delta}} - \frac{f(0)B}{\bar{\delta}} < z_{\alpha_i/2}\right) = o(1).$$

Recognizing the asymptotic normality of $\bar{S}_i(0)$ yields

$$\lim_{n\to\infty} P\left(\frac{\sqrt{2n}\bar{S}_i(B/\sqrt{2n})}{\bar{\delta}} < z_{\alpha_i/2}\right) = \Phi(z_{\alpha_i/2} + Bf(0)/\bar{\delta}).$$

Since $f(0)/\bar{\delta} > 0$ there exists a large enough B to ensure $\Phi(z_{\alpha_i/2} + Bf(0)/\bar{\delta}) > 1 - \epsilon/2$. We then have that

$$\liminf_{n\to\infty} P(\sqrt{2n}U_i \leq B) > 1 - \epsilon/2.$$

A similar argument shows that

$$\liminf_{n\to\infty} P(-\sqrt{2n}L_i \leq B) > 1 - \epsilon/2.$$

Noting that $L_i \leq U_i$ establishes both (i) and (ii).

PROOF OF THEOREM 4.1. Let $\boldsymbol{U_N} = \sqrt{N}(\bar{S}_1(0), \ldots, \bar{S}_m(0))^T$. As $N \to \infty$, $\boldsymbol{U_N}$ converges to an m-dimensional multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{D}$, with $d_i = \bar{\delta}^2/\lambda_i$, $i = 1, \ldots, m$ and $\operatorname{cov}(d_i, d_j) = \delta_{[1,2]}/\lambda_1$, $i \neq j$.

Let $\boldsymbol{A}$ be the $(m-1)$ by m matrix

$$\boldsymbol{A} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ -1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Then $\boldsymbol{A}\boldsymbol{U}_N$ converges to an $(m-1)$-dimensional multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, where $\sigma_i^2 = \bar{\delta}^2/\lambda_1 + \bar{\delta}^2/\lambda_i - 2\delta_{[1,2]}/\lambda_1$, $i = 2, \ldots, m$, and $\sigma_{i,j} = \bar{\delta}^2/\lambda_1 - \delta_{[1,2]}/\lambda_1$, $i \neq j$. The proof of the theorem follows from the definition of $\boldsymbol{T}$.

PROOF OF LEMMA 4.1. Note that $T^* = \max(T_2, \ldots, T_m)$ and the asymptotic distribution of $(T_2, \ldots, T_m)$ is normal with mean $\mathbf{0}$, variance and covariances

$$\operatorname{var}(T_i) = 1, \qquad \operatorname{cov}(T_i, T_j) = r = \frac{\bar{\delta}^2 - 2\delta_{[1,2]}}{m\bar{\delta}^2 - 2\delta_{[1,2]}}, \qquad i \neq j.$$

Let X_0 be a normal random variable with mean zero and variance r and $X_1, \ldots, X_{m-1}$ be iid normal random variable with mean zero and variance $1-r$. Assume that X_0 and $X_1, \ldots, X_{m-1}$ are independent. Then $(X_0 + X_1, \ldots, X_0 + X_{m-1})$ and $(T_2, \ldots, T_m)$ are equal in distribution. For the u-th quantile of T^*, we need to evaluate γ_u, such that

$$\begin{aligned} u &= P(T^* < \gamma_u) = P\left(\max_{1 \le i \le m-1}(|X_0 + X_1|, \ldots, |X_0 + X_{m-1}|) < \gamma_u\right) \\ &= \int P\left(\max_{1 \le i \le m-1}(|y + X_1|, \ldots, |y + X_{m-1}|) < \gamma_u \mid X_0 = y\right) f_{X_0}(y)dy \\ &= \int F^*_{m-1|y}(\gamma_u) f_{X_0}(y)dy, \end{aligned} \tag{A.1}$$

where $F^*_{m-1|y}(a)$ is the conditional cdf of the $\max(|y+X_1|, \ldots, |y+X_{m-1}|)$ given $X_0 = y$. It is clear that

$$F^*_{m-1|y}(\gamma_u) = \Phi^{m-1}\left(\frac{\gamma_u - y}{\sqrt{1-r}}\right) - \Phi^{m-1}\left(-\frac{\gamma_u + y}{\sqrt{1-r}}\right).$$

The proof is completed by putting this in equation (A.1).

PROOF OF THEOREM 4.2. From the definition of w_N we write that $\sqrt{N}w_N = \sqrt{N}(U_1 - L_1 + U_2 - L_2)$. Hence, from the linear representations in equations (3.2) and (3.3), it follows that

$$\begin{aligned} \sqrt{N}w_N &= \sqrt{N}(U_1 - L_1) + \sqrt{N}(U_2 - L_2) \\ &\to \frac{4\sqrt{\bar{\delta}^2 - \delta_{[1,2]}}\gamma_{\alpha_E}}{f(0)}. \end{aligned}$$

REFERENCES

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

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

Bhoj, D. S. (1997). New parametric ranked set sampling, *Journal of Applied Statistical Sciences*, **6**, 275–289.

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

Bohn, L. L. and 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.

Chakraborti, S. and Hettmansperger, T. P. (1996). Multi-sample inference for the simple tree alternative based on one-sample confidence interval, *Communications in Statistics. Theory and Methods*, **25**, 2819–2837.

Chen, Z. (2001). The optimal ranked-set sampling scheme for inference on population quantiles, *Statistica Sinica*, **11**, 23–27.

Dell, T. R. and Clutter, J. L. (1972). Ranked-set sampling theory with the use of ranked-set sampling on grass clover swards, *Grass and Forage Science*, **40**, 257–263.

Hettmansperger, T. P. (1984). Two-sample inference on one-sample sign statistics, *Journal of the Royal Statistical Society. Series C*, **33**, 45–51.

Hettmansperger, T. P. (1995). The ranked-set sample sign test, *Journal of Nonparametric Statistics*, **4**, 263–270.

Kaur, A., Patil, G. P., Sinha, A. K. and Taillie, C. (1995). Ranked set sampling: An annotated bibliography, *Environmental and Ecological Statistics*, **2**, 25–54.

Kaur, A., Patil, G. P. and Taillie, C. (1997). Unequal allocation models for ranked-set sampling with skew distributions, *Biometrics*, **53**, 123–130.

Koti, M. K. and Babu, G. J. (1996). Sign test for ranked-set sampling, *Communications in Statistics. Theory and Methods*, **25**, 1617–1630.

Kvam, P. H. and Samaniego, F. J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples, *Journal of the American Statistical Association*, **89**, 526–537.

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

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked-set sampling, *Australian Journal of Agricultural Research*, **3**, 385–390.

Nahhas, R. (1999). Ranked set sampling: Ranking error models, cost, and optimal set size, Ph.D. Thesis, Department of Statistics, The Ohio State University.

Nahhas, R., Wolfe, D. A. and Chen, H. Y. (2002). Ranked set sampling: Cost and optimal set size, *Biometrics*, **58**, 964–971.

Ozturk, O. (1999*a*). Two-sample inference based on one-sample ranked-set sample sign statistics, *Journal of Nonparametric Statistics*, **10**, 197–212.

Ozturk, O. (1999*b*). One and two-sample sign tests for ranked-set samples with selective designs, *Communications in Statistics, Theory and Methods*, **28**, 1231–1245.

Ozturk, O. (2002). Ranked-set sample inference under a symmetry restriction, *Journal of Statistical Planning and Inference*, **102**, 317–336.

Ozturk, O. and Wolfe, D. A. (2000*a*). Optimal allocation procedure in ranked set two-sample median test, *Journal of Nonparametric Statistics*, **13**, 57–76.

Ozturk, O. and Wolfe, D. A. (2000*b*). An improved ranked-set two-sample Mann-Whitney-Wilcoxon test, *Canadian Journal of Statistics*, **28**, 123–135.

Ozturk, O. and Wolfe, D. A. (2000*c*). Optimal allocation procedure in ranked-set sampling for unimodal and multi-modal distributions, *Environmental and Ecological Statistics*, **7**, 343–356.

Ozturk, O. and Wolfe, D. A. (2000*d*). Alternative ranked set sampling protocols for the sign test, *Statistics and Probability Letters*, **47**, 15–23.

Ozturk, O., Wolfe, D. A. and Alexandridis, R. (2004). Multi-sample inference for simple-tree alternatives with ranked-set samples, *Australian and New Zealand Journal of Statistics*, **46**, 443–455.

Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*, 2nd ed., Duxbury, Pacific Grove, California.

Randles, H. R. and Wolfe, D. A. (1991). *Introduction to the Theory of Nonparametric Statistics*, Krieger Publishing Company, Malabar, Florida.

Sinha, B. K., Sinha, B. K. and Purkayastha, S. (1996). On some aspects of ranked set sampling for estimation of normal and exponential parameters, *Statistics and Decisions*, **14**, 223–240.

Stark, G. (2001). Imperfect ranking models and their use in the evaluation of ranked-set sampling procedure, Ph.D. Thesis, Department of Statistics, The Ohio State University.

Stokes, S. L. (1977). Ranked set sampling with concomitant variables, *Communications in Statistics. Theory and Methods*, **12**, 1207–1211.

Stokes, S. L. (1980). Estimation of variance using judgment ordered ranked set samples, *Biometrics*, **36**, 35–42.

Stokes, S. L. and Sager, T. W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions, *Journal of the American Statistical Association*, **83**, 374–381.

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

Yu, P. L., Lam, K. and Sinha, B. K. (1999). Estimation of normal variance based on balanced and unbalanced ranked set samples, *Environmental and Ecological Statistics*, **6**, 23–46.