Sankhyā : The Indian Journal of Statistics
Special Issue on Quantile Regression and Related Methods
2005, Volume 67, Part 2, pp 295-304

Quantile Estimation from Ranked Set Sampling Data

Min Zhu
CSIRO Mathematical and Information Sciences, Australia
You-Gan Wang
National University of Singapore and CSIRO, Floreat Park, Australia

Abstract

We consider estimation of quantiles when data are generated from ranked set sampling. A new estimator is proposed and is shown to have a smaller asymptotic variance for all distributions. It is also shown that the optimal sampling strategy is to select observations with one fixed rank from different ranked sets. Both the optimal rank and the relative efficiency gain with respect to simple random sampling are distribution-free and depend on the set size and the given probability only. In the case of median estimation, it is analytically shown that the optimal design is to select the median from each ranked set.

AMS (2000) *subject classification.* 62G32, 62G99.
Keywords and phrases. Asymptotic variance, efficiency, optimal design, ranked set sampling, quantile estimation.

1 Introduction

Rank set sampling (RSS) was first proposed by McIntyre (1952) for estimating the mean pasture yields. Since then there has been substantial progress in studying this sampling scheme and statistical inference from RSS data. We now briefly introduce the concept of RSS for completeness.

Suppose X is a random variable with a density function $f(x)$. If $(X_1, X_2, ..., X_k)$ are the unobserved values from k units, we may rank them by visual inspection or based on a concomitant variable. Ranked set sampling (RSS) involves selecting one unit among every ranked set consisting of k units for quantification, with the other $k - 1$ units not being investigated further. One may select the unit with rank 1 from the first set, and rank 2 from the second set, and so on. The first cycle is completed when the unit with rank

k is selected from the k-th set. Of course, the selected rank order can be any permutation of $1, 2, ...k$. Each cycle involves k^2 units and among which only k units will be selected for quantification. This cycle can be repeated for a certain number of times. This sampling scheme has a balanced nature in the sense that equal number of observations will be obtained from each rank. An important property is that the sample mean is an unbiased estimator of the population mean regardless of whether ranking is perfect or not (Dell and Clutter, 1972). It can be easily shown that the sample mean using RSS has a smaller variance than the sample mean using the traditional simple random sampling (SRS) when the number of observations are the same. Clearly, RSS may be considered only when taking measurements on units are difficult or costly while ranking a set of the units is easy.

Many contributions have been made towards nonparametric inference from RSS data (see, for example, Öztürk, 1999; Bohn and Wolfe, 1992; Hettmansperger, 1995). Stokes and Sager (1988) first considered estimation of distribution function from balanced RSS data. This work was generalized nicely by Kvam and Samaniego (1994). In some cases, an unbalanced RSS design may further improve statistical efficiency. For example, for one sample sign tests, the Pitman efficiency can achieve its maximum when selecting the medians from each ranked set (Öztürk, 1999; Wang and Zhu, 2005). Regression estimators from RSS data have also been investigated (Yu and Lam, 1997; Barnett and Moore, 1997; Chen and Wang, 2004). Wang et al. (2004) also considered optimal designs for general ranked set sampling when more than one unit can be selected from each ranked set.

Chen (2000) first considered quantile estimation from balanced RSS data and found that the RSS method can substantially improve the efficiency of the quantile estimators. This can be regarded as a generalization to Hettmansperger (1995) who considered sign tests. In a follow-up paper, Chen (2001) further generalized his results to unbalanced designs in which the proportion of observations associated with rank r, q_r, can be any value satisfying $0 \leq q_r \leq 1$ and $\sum_{r=1}^{k} q_r = 1$. The quantile estimator considered by Chen (2000, 2001) is based on the empirical distribution of the pooled RSS data as in Stokes and Sager (1988) who considered the balanced cases.

In this paper, we will consider quantile estimation when ranking is perfect. We will first propose a new quantile estimator which is motivated by the fact that the observations from RSS are not identically distributed. In other words, observations selected by different ranks should contribute differently because they follow different distributions. It is proven analytically that the new estimator has a smaller asymptotic variance in general including the balanced design. We then consider how to select the allocation sequence

(q_j) to minimize the asymptotic variance. It is shown that the optimal design can rely on selecting observations with one fixed rank. In particular, for median estimation, we prove analytically that the optimal design is to select the median from each ranked set. We also discussed briefly how one may construct an analogous estimator for distribution functions.

2 Quantile Estimation

Suppose the cumulative distribution of the population is $F(x)$ and we wish to estimate the pth quantile, $\xi_p = \inf\{x : F(x) \geq p\}$. For convenience, we denote the Beta density function $u^{a-1}(1-u)^{b-1}/B(a,b)$ as $g(x;a,b)$, and the cumulative distribution function as $G(x;a,b)$, where $0 \leq x \leq 1$ and $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}du$.

We assume that the number of observations associated with rank j is n_j, and the data consist of independent observations $(x_{[j]i})$, where $1 \leq j \leq k$ and $1 \leq i \leq n_j$. The balanced design is a special case when $n_j = n$ for all $j = 1, 2, ..., k$. The selective designs are also special cases here. Let $N = \sum_{j=1}^k n_j$, the total number of observations and $q_j = \lim_{N\to\infty} n_j/N$, the asymptotic proportion of observations with rank j. To simplify the algebra, we will assume $q_j = n_j/N$ and denote the vector (q_j) as $\mathbf{q}$. Here we assume that the proportion vector $\mathbf{q}$ is independent of the observations. We first consider the case when $\mathbf{q}$ is known and data are already collected. Later on we will consider how to choose $\mathbf{q}$ to minimize the variance of the estimator.

For each j, $x_{[j]1}, x_{[j]2}, ...,$ are independent realizations of jth order statistic that has the density function $f_{[j]}(x) = g(F(x); j, k+1-j)f(x)$. The cumulative distribution function is

$$F_{[j]}(x) = G(F(x); j, k+1-j), \tag{2.1}$$

which implies

$$F_{[j]}(\xi_p) = G(p; j, k-j+1). \tag{2.2}$$

Therefore, ξ_p can be estimated by the quantile of $F_{[j]}$ corresponding to the probability of $G(p; j, k-j+1)$, which is a known constant for any given p. Hettmansperger (1995) also applied this relationship in the context of sign tests. Let $\hat{F}_{[j]}(x)$ be the empirical distribution for $F_{[j]}(x)$ using observations with rank j only. For each j, when $q_j > 0$, we can obtain an estimator for ξ_p based on the observations with rank j

$$\hat{\xi}_{[j]} = \inf\{x : \hat{F}_{[j]}(x) \geq G(p; j, k+1-j)\}.$$

For convenience, we will denote $G(p;j,k+1-j)$ as G_j and $g(p;j,k+1-j)$ as g_j when there is no confusion. From Chen (2001), we know that, each $\hat{\xi}_{[j]}$ is a consistent estimator for ξ_p and has asymptotic variance w_j^{-1}, where $w_j = n_j g_j^2/\{G_j(1-G_j)\}$. Note that even in the balanced case, i.e. $n_j = n$, w_j is not a constant. A reasonable way is to combine them and obtain a weighted estimator for ξ_p,

$$\hat{\xi}_W = \left(\sum_{j=1}^{k} w_j\right)^{-1} \left(\sum_{j=1}^{k} w_j \hat{\xi}_{[j]}\right), \tag{2.3}$$

where w_j is the inverse of the asymptotic variance of $\hat{\xi}_{[j]}$. The asymptotic variance of $\sqrt{N}\hat{\xi}_W$ is

$$V_W = \left\{\sum_{j=1}^{k} \frac{q_j f_{[j]}^2(\xi_p)}{G_j(1-G_j)}\right\}^{-1}.$$

Chen (2001) suggested to consider the mixture distribution $F_{\mathbf{q}} = \sum_{j=1}^{k} q_j F_{[j]}(x)$ of order statistics. This is appropriate when we ignore the rank information associated with each observation or when their ranks are not known. The empirical distribution of $F_{\mathbf{q}}(x)$ is $\sum_{j=1}^{k} q_j \hat{F}_{[j]}(x)$. However, we are interested in estimating ξ_p, the pth quantile of $F(x)$, instead of $F_{[j]}(x)$ or $F_{\mathbf{q}}(x)$. This quantile ξ_p for $F_{\mathbf{q}}()$ corresponds to the cumulative probability of

$$F_{\mathbf{q}}(\xi_p) = \sum_{j=1}^{k} q_j G(p;j,k+1-j),$$

a known constant. This implies that ξ_p is the sth quantile of $F_{\mathbf{q}}(x)$, where $s = \sum_{j=1}^{k} q_j G(p;j,k+1-j)$. This leads to the following estimator for ξ_p,

$$\hat{\xi}_C = \inf\{x : \hat{F}_{\mathbf{q}}(x) \geq \sum_{j=1}^{k} q_j G(p;j,k+1-j)\}.$$

The asymptotic variance of $\sqrt{N}\hat{\xi}_C$ is

$$V_C = \frac{\sum_{j=1}^{k} q_j G(p;j,k+1-j)\{1-G(p;j,k+1-j)\}}{\{\sum_{j=1}^{k} q_j f_{[j]}(\xi_p)\}^2}. \tag{2.4}$$

The following results (2.5), (2.6) and (2.7) can be easily obtained following Theorem 1 and 2 in Chen (2001).

THEOREM 1. *Under the assumption that $f(x)$ is positive and continuous at ξ_p, we have*

$$|\hat{\xi}_W - \xi_p| \leq \max_{1\leq j\leq k} \frac{2(\log n_j)^2}{n_j^{1/2} f_{[j]}(\xi_p)} \text{ with probability } 1, \tag{2.5}$$

$$\hat{\xi}_W - \xi_p = \frac{\sum_j q_j g_j \{G_j - \hat{F}_{[j]}(\xi_p)\}}{f(\xi_p) \sum_j q_j g_j^2 / \{G_j(1-G_j)\}} + O\left(N^{-3/4} (\log N)^{3/4}\right) \tag{2.6}$$

and

$$\sqrt{N}(\hat{\xi}_W - \xi_p) \to N(0, V_W) \text{ in distribution.} \tag{2.7}$$

The estimator $\hat{\xi}_C$ is based on a mixture distribution of order statistics. This is reasonable if we do not know which rank group of each observation is associated. However, for each observation, we do know the associated rank. Therefore, it should be statistically more efficient to make use of this information, i.e. we should regard each observation is generated from $f_{[j]}(x)$ where j is its associated rank instead of $f_{\mathbf{q}}(x)$. This motivates the new estimator, $\hat{\xi}_W$. According to the Cauchy-Schwarz inequality, we can easily show that $V_W^{-1} \geq V_C^{-1}$, and hence we have $V_W \leq V_C$. This shows that the new estimator always has a smaller variance.

The case of balanced design is of special interest. The balanced case is also considered by Stokes and Sager (1988) for estimating $F(x)$. Their method is also based on the empirical distribution of the pooled data as Chen (2001). In fact, when $k \geq 3$, we always have $V_W < V_C$. Table 1 shows the relative efficiency of V_C and V_W with respect to the simple random sampling (SRS) for $k = 3, 4, ..., 20$ and $p = 0.01, 0.05, 0.1, 0.25$ and 0.5. Interestingly, the efficiency gain in general increase in k and as well as p. The gain over the SRS is quite substantial. The new estimator always outperforms $\hat{\xi}_C$ slightly.

3 Optimal Sampling Scheme

We have considered quantile estimation for RSS data. We now consider the design problem – how to choose (q_j) values so that the asymptotic variance of the estimator is minimized. The following Theorem shows how to choose the design sequence (ξ_j) to minimize V_W. We state our conclusion in the following theorem.

THEOREM 2. *There exists an integer τ, between 1 and k so that the minimum asymptotic variance of $\hat{\xi}_W$ is achieved by letting $q_j = 1$ for $j = \tau$*

TABLE 1. COMPARISON OF THE ASYMPTOTICAL RELATIVE EFFICIENCY (TO THE SIMPLE RANDOM SAMPLING) OF THE QUANTILE ESTIMATOR BY CHEN (2001) ($\hat{\xi}_C$) AND THE NEW ESTIMATOR ($\hat{\xi}_W$) AT DIFFERENT PROBABILITY VALUES (p) FOR BALANCED DESIGNS. THESE RESULTS ARE INDEPENDENT OF THE DISTRIBUTION OF OBSERVATIONS.

k	$p = 0.25$		$p = 0.25$		$p = 0.25$		$p = 0.25$		$p = 0.25$	
	$\hat{\xi}_C$	$\hat{\xi}_W$	$\hat{\xi}_C$	$\hat{\xi}_W$	$\hat{\xi}_C$	$\hat{\xi}_W$	$\hat{\xi}_C$	$\hat{\xi}_W$	$\hat{\xi}_C$	$\hat{\xi}_W$
3	1.020	1.029	1.099	1.135	1.196	1.245	1.438	1.473	1.600	1.607
4	1.030	1.044	1.149	1.197	1.290	1.352	1.625	1.663	1.829	1.842
5	1.040	1.058	1.198	1.256	1.382	1.451	1.794	1.835	2.032	2.051
6	1.050	1.072	1.246	1.312	1.471	1.544	1.950	1.993	2.216	2.241
7	1.060	1.086	1.294	1.367	1.557	1.633	2.095	2.140	2.387	2.416
8	1.070	1.099	1.341	1.419	1.639	1.717	2.230	2.278	2.546	2.579
9	1.080	1.113	1.387	1.470	1.719	1.798	2.358	2.408	2.696	2.733
10	1.090	1.126	1.433	1.519	1.796	1.876	2.480	2.531	2.838	2.878
11	1.100	1.139	1.478	1.566	1.871	1.951	2.595	2.649	2.973	3.016
12	1.110	1.152	1.523	1.613	1.943	2.024	2.706	2.762	3.102	3.149
13	1.120	1.165	1.566	1.658	2.013	2.094	2.813	2.870	3.226	3.276
14	1.130	1.178	1.609	1.702	2.081	2.162	2.916	2.975	3.346	3.398
15	1.139	1.191	1.652	1.746	2.147	2.228	3.015	3.076	3.461	3.516
16	1.149	1.203	1.693	1.788	2.211	2.292	3.111	3.174	3.573	3.630
17	1.159	1.216	1.734	1.829	2.273	2.355	3.204	3.269	3.681	3.741
18	1.169	1.228	1.774	1.870	2.334	2.416	3.295	3.361	3.786	3.848
19	1.179	1.240	1.814	1.910	2.393	2.475	3.383	3.451	3.888	3.953
20	1.189	1.252	1.853	1.949	2.451	2.533	3.469	3.538	3.988	4.054

and 0 otherwise, i.e. selecting all the samples with rank τ from each ranked set. In particular, for median estimation, we have $\tau = (k+1)/2$ if k is odd, and τ can take either $k/2$ or $k/2+1$ if k is even.

PROOF. For convenience, define

$$d_j = \frac{g^2(p; j, k+1-j)}{G(p; j, k+1-j)\{1 - G(p; j, k+1-j)\}}.$$

For any nonnegative sequence (q_j) satisfying $\sum_{j=1}^{k} q_j = 1$, we have

$$V_C \geq V_W \geq \min_{1\leq j\leq k} d_j^{-1} = d_\tau^{-1},$$

where $\tau = \text{argmin}_{1\leq j\leq k} d_j^{-1}$. In the case of multiple js achieving the minimum, we will select the smallest among them for τ. Clearly, when $q_j = 1$ for $j = \tau$ and 0 for $j \neq \tau$, we have $V_C = V_W = d_\tau^{-1}$, achieving the minimum asymptotic variance. This completes the first part of the Theorem.

For median estimation ($p = 0.5$), we now show that the optimal τ is the median rank. Define τ as $(k+1)/2$ if k is odd and $k/2$ (or $k/2+1$) if k is even. It suffices to show that, for any j between 1 and k, $d_j^{-1} \geq d_\tau^{-1}$, i.e.

$$\frac{G(0.5; j, k{+}1{-}j)G(0.5; k+1-j, j)}{g_j^2} \geq \frac{G(0.5; \tau, k+1-\tau)G(0.5; k{+}1{-}\tau, \tau)}{g_\tau^2}.$$

Since $g_j = g_\tau$ when $p = 0.5$, the above expression is equivalent to

$$\int_0^{0.5} x^{j-1}(1{-}x)^{k-j} dx \int_0^{0.5} x^{k-j}(1{-}x)^{j-1} dx \geq \int_0^{0.5} x^{\tau-1}(1{-}x)^{k-\tau} dx \int_0^{0.5} x^{k-\tau}(1{-}x)^{\tau-1} dx.$$

For odd k, we have by the Cauchy-Schwarz inequality,

$$\int_0^{1/2} x^{j-1}(1{-}x)^{k-j} dx \int_0^{1/2} x^{k-j}(1{-}x)^{j-1} dx \geq \left\{ \int_0^{1/2} x^{(k-1)/2}(1{-}x)^{(k-1)/2} dx \right\}^2,$$

and equality holds when $j = (k+1)/2 = \tau$. For even k, $\tau = k/2$, define $p = 2-1/(\tau-j+1)$ and $q = 2+1/(\tau-j)$ satisfying $1/p+1/q = 1$. Applying the Holder inequality, we have

$$\begin{aligned} \int_0^{1/2} x^{\tau-1}(1{-}x)^{\tau} dx &= \int_0^{1/2} x^{j-1}(1-x)^{k-j}\}^{1/p} \left\{ x^{k-j}(1-x)^{j-1} \right\}^{1/q} dx \\ &= \left\{ \int_0^{1/2} x^{j-1}(1{-}x)^{k-j} dx \right\}^{1/p} \left\{ \int_0^{1/2} x^{k-j}(1{-}x)^{j-1} dx \right\}^{1/q}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \int_0^{1/2} x^{\tau}(1{-}x)^{\tau-1} dx &= \int_0^{1/2} \{x^{j-1}(1{-}x)^{k-j}\}^{1/q} \left\{ x^{k-j}(1{-}x)^{j-1} \right\}^{1/p} dx \\ &\leq \left\{ \int_0^{1/2} x^{j-1}(1{-}x)^{k-j} dx \right\}^{1/q} \left\{ \int_0^{1/2} x^{k-j}(1{-}x)^{j-1} dx \right\}^{1/p}. \end{aligned}$$

We therefore have for an even k,

$$\int_0^{1/2} x^{\tau-1}(1-x)^{\tau}dx \int_0^{1/2} x^{\tau}(1-x)^{\tau-1}dx \leq \int_0^{1/2} x^{j-1}(1-x)^{k-j}dx \int_0^{1/2} x^{k-j}(1-x)^{j-1}dx,$$

and the equality clearly holds when $j = \tau$. When selecting only the median observations, i.e. $q_\tau = 1$ and $q_j = 0$ for all $j \neq \tau$, the minimum variance of $\hat{\xi}_W$ is achieved. Note that because $V_C \geq V_W$, selecting the medians will also achieve the minimum variance of $\hat{\xi}_W$ as well as for $\hat{\xi}_C$. Chen (2001) concluded that the optimal $\mathbf{q}$ will have at most two non-zero components. Our Theorem 2 here implies that the optimal $\mathbf{q}$ for $\hat{\xi}_C$ can only consist of one non-zero component, namely, $q_\tau = 1$. Our analytic results are also supported by the numerical findings in Chen (2001). When only one rank is used, observations will consist of i.i.d. samples from the distribution $F_{[\tau]}(x)$. The two estimators $\hat{\xi}_W$ and $\hat{\xi}_C$ will coincide when a single rank is used. The relative efficiency with respect to the simple random sampling can be found in Chen (2001). Because our new estimator is more efficient than $\hat{\xi}_C$, the design using a single rank τ is therefore also an optimal design for the approach of Chen (2001).

For median estimation, according to Theorem 2, selecting the medians from each ranked set is optimal. Note that the optimal $\mathbf{q}$ may not be unique. For example, when k is even, any allocation sequence (q_j) satisfying $q_\tau + q_{\tau+1} = 1$ will also be optimal.

The proposed approach is also applicable to estimation of multiple quantiles. If $\hat{\xi}_{[j]}$ is a vector estimator for the quantiles $(p_1, p_2, ...p_s)$ based on $\hat{F}_{[j]}$, and $w_j^{-1} = \text{var}\{\hat{\xi}_{[j]}\}$. The expression given by (2.3) is still applicable to obtain the combined estimator $\hat{\xi}_W$. Intuitively, this estimator should always be more efficient than $\hat{\xi}_C$. An analytic proof is still not available. Also the optimal sampling scheme will involve multiple ranks instead of one, and we will have to rely on numerical approach to find the solution.

4 Discussion

A very related problem to quantile estimation is estimation of distribution functions. Stokes and Sager (1988) showed that the RSS can substantially improve the efficiency in estimating $F(x)$ (for a given x). An analogue of the weighted quantile estimator may be developed for this problem.

For each j, if $q_j > 0$, (2.1) implies that we can obtain a consistent estimator of $F(x)$ from observations with rank j as,

$$\hat{F}_j^*(x) = H(\hat{F}_{[j]}(x); j, k+1-j),$$

in which H is the quantile function for the Beta distribution $G(., j, k+1-j)$ and $\hat{F}_{[j]}(x)$ is the empirical distribution of $F_{[j]}$ using data with rank j. We may weight $\hat{F}_j^*(x)$ by the inverse of its asymptotic variance to obtain a pooled estimator for $F(x)$. It can be shown that the optimal allocation sequence is to select observations with rank j so that $j/(k+1) \approx F(x)$. However, to apply this design, one need to have some priori knowledge on $F(x)$. If we are interested in estimating the function curve $F(x)$ instead of at a given x, the balanced design would be a reasonable choice. Further research in this direction may be carried out to evaluate its performance and compare with the nonparametric maximum likelihood estimator by Kvam and Samaniego (1994).

We have focused on quantile estimation and proposed a new estimator for the pth quantile which is shown to be more efficient asymptotically when ranking is perfect. It would be interesting to see how these estimators perform in practice when ranking is imperfect. If the probability distribution of ranking can be specified we can also obtain a consistent estimator and its standard error (Chen 2001). A modified estimator similar to $\hat{\xi}_W$ can also be constructed. However, the efficiency gain will deteriorate as ranking is prone to error. Therefore, further improvement in efficiency may be of little practical value especially when ranking is not so accurate. Also when ranking is imperfect, the optimal scheme is sensitive to the assumed probabilities of ranking errors and it also depends on distribution of ranking errors (Chen, 2001). These are the major disadvantages of the RSS method for quantile estimation. However, imperfect ranking is the norm in practice. It is therefore of great importance to carry out further research to better understand the implications of imperfect ranking.

Selective designs have been considered by a few authors in their numerical studies (see Muttlak, 2001; Öztürk and Wolfe, 2000a, 2000b; Chen 2001). However, no analytical results are available in the literature. Our results on the optimal allocation sequence provide a theoretical justification in using selective designs for quantile estimation. This should be of theoretical significance as well as of practical importance.

Acknowledgments. We are grateful to a referee who provided very constructive comments.

References

Barnett, V. and Moore, K. (1997). Best linear unbiased estimates in ranked-set sampling with particular reference to imperfect ordering. *J. Appl. Statist.*, **24**, 697-710.

Bohn, L.L. and Wolfe, D.A. (1992). Nonparametric two-sample procedures for ranked-set samples data. *J. Amer. Statist. Assoc.*, **87**, 552-561.

Chen, Z. (2000). On ranked-set sample quantiles and their applications. *J. Statist. Plann. Inference*, **83**, 125-135.

Chen, Z. (2001). The optimal ranked-set sampling scheme for inference on population quantiles. *Statist. Sinica*, **11**, 23-37.

Chen, Z. and Wang, Y.-G. (2004). Efficient regression analysis with ranked set sampling. *Biometrics*, **60**, 997-1004.

Dell, T. R. and Clutter, J.L. (1972). Ranked set sampling theory with order statistics background. *Biometrika*, **28**, 545-555.

Hettmansperger, T.P. (1995). The ranked-set sample sign test. *J. Nonparametr. Statist.*, **4**, 263-270.

Kvam, P.H. and Samaniego, F.J. (1994). Nonparametric maximum likelihood estimation based on ranked set samplings. *J. Amer. Statist. Assoc.*, **89**, 526-537.

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

Muttlak, H.A. (2001). Regression estimators in extreme and median ranked set samples. *J. Appl. Statist.*, **28**, 1003-1017.

Öztürk, O. (1999). Two sample inference based on one sample ranked set sample sign statistics. *J. Nonparametr. Statist.*, **10**, 197-212.

Öztürk, O. and Wolfe, D.A. (2000a). Alternative ranked set sampling protocols for the sign test. *Statist. Probab. Lett.*, **47**, 15–23.

Öztürk, O. and Wolfe, D.A. (2000b). An improved ranked set two-sample Mann-Whitney-Wilcoxon test. *Canad. J. Statist.* **28**, 123-135.

Stokes, S.L. and Sager, T.W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *J. Amer. Statist. Assoc.*, **83**, 374-381.

Wang, Y.-G., Chen, Z. and Liu, J. (2004). General ranked set sampling with cost considerations. *Biometrics* **60**, 556-561.

Wang, Y.-G. and Zhu, M. (2005). Efficient sign tests from ranked set sampling data. *Statist. Prob. Letters* **72**, 13-22.

Yu, P. L. H. and Lam, K. (1997). Regression estimator in ranked set sampling. *Biometrics*, **53**, 1070-1080.

Min Zhu and You-Gan Wang
CSIRO Mathematical and Information Sciences
65 Brockway Road
Floreat, WA 6014, Australia
E-mail: Min.Zhu@csiro.au
You-Gan.Wang@csiro.au

Paper received: February 2004; revised March 2005.