# On Ranked Set Sampling for Multiple Characteristics

M.S. Ridout

Institute of Mathematics and Statistics
University of Kent, Canterbury
Kent CT2 7NF U.K.

**Abstract**

We consider the selection of samples in ranked set sampling when several attributes of each sample are of interest. We describe approaches that have appeared previously in the literature and present a novel method that seeks to achieve samples that are nearly balanced with respect to the ranks of all attributes. This method is shown to result in very little loss of precision compared to problems in which only a single sample attribute is of interest.

*Keywords*: balanced allocation, concomitant variable, multiple attributes, relative precision.

## 1 Introduction

Ranked set sampling is an established sampling technique that has been applied particularly in agricultural and environmental sciences. For an overview, see, for example, Patil, *et al* (1994a). In its simplest form, known as *balanced* ranked set sampling, the procedure is as follows. Random samples of size $m$ are selected from the population and ranked by some means other than direct measurement, for example visually or using a concomitant measurement that is cheaper than the measurement of interest. One sample from each set is measured precisely; the sample ranked smallest is chosen from the first set, the sample ranked second smallest is chosen from the second set and so on, until the sample ranked largest is chosen from the $m$th set. This completes one *cycle* of the procedure and, in general, this is repeated $r$ times. Thus, altogether, $rm^2$ samples are ranked, but only $n = rm$ of these are measured.

Ranked set sampling improves the precision with which the population mean is estimated, relative to simple random sampling with the same sample size, $n$. This is true even if there are errors in ranking, although the relative precision decreases as ranking errors increase, and ranked set sampling is equivalent to random sampling if ranking is no better than random. In practice, the key issue is whether the increase in precision is sufficient to justify the increased costs associated with the ranking process. When the population distribution is highly skewed, as is often the case for environmental variables, the precision of ranked set sampling may be increased by using an *unbalanced allocation*, with more high-ranked samples selected for measurement (Kaur, *et al*, 1997, Barnett, 1999). More generally, in the absence of ranking errors, the optimal allocation, in the sense of minimising the variance of the estimator of the population mean, is the Neyman allocation, in which the number of samples measured at the $j$th rank is proportional to the standard deviation of the $j$th order statistic (Takahasi and Wakimoto, 1968).

In some situations, we may be interested in measuring several attributes of each sample. In the particular example that motivated this work the sampling unit was a leaf and the amounts of spray deposited on both the upper and lower leaf surfaces were of interest. Patil, *et al* (1994b) discuss this problem. Following McIntyre (1952), they assume that one attibute is of primary interest and base the selection of samples on the ranking with respect to this attribute, which therefore acts as a concomitant variable for the other attributes. This works reasonably well if the primary attribute is highly correlated with the other attributes of interest, since the rankings of different attributes are then likely to be similar, but is unsatisfactory for the leaf samples, because there is only a weak negative correlation between the spray deposits on the upper and lower surfaces.

In a subsequent paper, Norris, *et al* (1995) discuss two alternative approaches. One is a modification of an idea of Takahasi (1970), which we discuss below. The second is to use an unbalanced allocation based on the Neyman allocation for the attribute of primary interest, treating this as a concomitant for the other variables of interest. We do not pursue this second approach here for two reasons. First, it will not work well unless the variables of interest are highly correlated. Second, it relies on knowledge of the standard deviations of the order statistics, and these are typically unknown. Norris, *et al* (1995) recognise this, and suggest that estimates might be obtained from a small pilot ranked set sampling study with equal allocation. However, such estimates have large sampling errors and can lead to allocations that are *less* efficient than balanced rank set sampling.

In this paper we compare several possible approaches that involve ranking the samples within a set separately for *each* attribute. Whilst this involves some extra effort, it can improve efficiency quite considerably; for example, if we select samples solely on the basis of the ranking of one attribute, we can nonetheless analyse other attributes using their own ranks and this will typically give a much more efficient estimator than using the ranks assigned to the primary attribute, as we show later.

The paper is organised as follows. The spray deposit example is introduced briefly in Section 2, to provide background and motivation. Section 3 describes several methods of selecting samples and Section 4 evaluates these using simulated data; one method is shown to be particularly useful. Section 5 concludes by discussing some practical aspects.

## 2 A motivating example

This example arose as part of an extensive practical evaluation of ranked set sampling for estimating spray deposits on the leaves of fruit trees. Since the main results of this study will be reported elsewhere, we provide only brief details here.

Apple trees were sprayed with a fluorescent dye, using one of two sprayer settings. For each setting, 125 leaves were collected haphazardly from the sprayed trees and taken to the laboratory where they were divided randomly into 25 sets of five. The upper and lower surfaces of the leaves in each set were then ranked separately under ultraviolet light by four rankers, working independently. Accurate measurements of deposit per unit area were made using an image analysis system. For this evaluation study, all samples were measured.

Estimates of relative precision compared to simple random sampling varied between sprayer settings, between leaf surfaces and between rankers. Estimates of relative precision would have been about 2.5, had ranking been perfect, but were about 2.1 when ranking errors were taken into account. When the additional costs of ranked set sampling were also taken into account, the estimated relative efficiency was only about 1.6 (range 1.45–1.75); however, this figure still represents a worthwhile improvement in efficiency compared to simple random sampling.

These estimates assume, however, that leaf surfaces are ranked and sampled separately. In practice one would wish to measure the upper and lower surfaces of the *same* leaves, partly to minimise costs, and partly because derived variables, such as the ratio of the deposits on the two surfaces, may be of interest. This presents a difficulty, because obtaining a balanced sample with respect to the ranks on one surface may give a very unbalanced sample with respect to the ranks on the other surface. For an unbalanced sample, with $r_j$ observations at rank $j$, the ranked set sampling estimator of the population mean is

$$\hat{\mu}_{RSS} = \frac{1}{m} \sum_{j=1}^{m} \frac{T_j}{r_j}, \tag{1}$$

where $T_j$ is the sum of the observations that were assigned rank $j$. In extreme cases, one or more of the $r_j$ might be zero, if samples are selected on the basis of the ranking of another attribute, and there is then no unbiased estimator of the population mean. Imbalance would be less of a problem if the deposits on the two surfaces were highly correlated, either positively or negatively, but the correlations for the two spraying methods were $-0.15$ and $-0.37$. The negative correlations arise because leaves that have one surface favourably orientated towards the sprayer typically receive a low deposit on the other surface. However, many leaves get a low deposit on both surfaces, and hence the correlations are only weak.

## 3 Methods of selecting samples

In this section we continue to use the terminology of the spray deposit example. We assume that rankings have been obtained for both leaf surfaces and consider four methods of selecting the samples that are to be measured. Extensions to more than two variables are considered in Section 4.

**Method 1** Select samples so as to give a balanced ranked set sample with respect to just one surface, say the upper surface. The other surface is ignored completely when selecting samples, but estimation for this surface is based on the (usually unbalanced) rankings of the sample that is obtained, using equation (1). The disadvantages of this method have already been discussed.

**Method 2** Since both surfaces are of equal interest, select samples for half of the cycles based on the upper surface ranking, and for the remainder use the lower surface ranking. This spreads any imbalance (and any resulting loss of efficiency) across both leaf surfaces, and in particular eliminates the possibility of failing to obtain any samples at a particular rank.

**Method 3** Select a sample *at random* from each set, but use the rank of the sample in estimation. This method was originally suggested, for a single attribute, by Takahasi (1970), who gave a formula for its precision relative to balanced ranked set sampling. Although random selection is less efficient, the relative precision tends to one as $n \to \infty$. Norris, *et al* (1995) have pointed out that the method may be useful for multiple attributes. A disadvantage of this method is that, as with method 1, it may fail to give samples at every rank. For a single attribute, Takahasi (1970) suggested using one preliminary cycle of balanced rank set sampling to avoid this problem and Norris, *et al* (1995) suggested extending this to two initial cycles, so that the variance of the ranked set sampling estimator can be estimated.

**Method 4** Select samples according to the following algorithm. Label the upper and lower leaf surfaces as $i = 1$ and $i = 2$, respectively. Let $n_{i[j]}$ denote the number of samples that have so far been assigned rank $j$ for leaf surface $i$, and let $V_i$ denote the sample variance of the $n_{i[j]}$. For a balanced sample, $V_i = 0$. Given a new set of size $m$, compute what the updated values of $V_1$ and $V_2$ would be for each of the $m$ samples in turn, and select the sample that minimises $V = V_1 + V_2$; if there is more than one such sample, select one of them at random.

An example may be useful to illustrate method 4. Suppose that we are collecting a ranked set sample of 25 leaves ($m = 5$, $r = 5$), and that the following distribution of ranks has been obtained from the first 23 sets:

| Rank, $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Upper surface | 5 | 5 | 4 | 4 | 5 |
| Lower surface | 5 | 4 | 5 | 5 | 4 |

The current value of $V$ is 0.6. The rankings of the next set of five leaves are as follows:

| Leaf number | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Upper surface rank | 2 | 1 | 3 | 4 | 5 |
| Lower surface rank | 5 | 1 | 3 | 2 | 4 |
| $V$ | 0.9 | 1.4 | 0.9 | 0.4 | 1.4 |

Method 4 would select the fourth leaf, because this minimises the updated value of $V$.

There is a link between methods 1 and 3 if the different attributes are independent. Under method 1, if ranking is based on the upper leaf surface deposit, each lower leaf surface measurement will effectively be a random sample from a set of $m$ leaves (because the measurements are independent). Method 3 involves random sampling for both attributes and so will achieve the same relative precision as Method 1 for the lower leaf surface, but a lower relative precision for the upper leaf surface. When the upper and lower surface measurements are correlated, the leaves sampled under method 1 will tend to be more balanced with respect to the lower leaf surface ranking than under independence. Thus there is no reason to use method 3 in preference to method 1 (the modification applied in method 3, of doing one cycle of ranked set sampling for each attribute, can also be adopted in method 1). We therefore limit our comparisons to methods 1, 2 and 4.

We can, moreover make use of Takahasi's (1970) result that relative precision under random selection is reduced in comparison to relative precision under balanced selection by a factor of

$$c = \frac{n-m+1}{n}\left[1-\left(1-\frac{1}{m}\right)^{n-m+1}\right]^{-1}. \tag{2}$$

In the present context, $c$ provides a lower bound for the factor by which relative precision is reduced in method 1 for the attribute whose ranking is *not* used for selecting samples. For method 2, a lower bound (for both attributes) is $(1+c)/2$.

Although method 4 is described in terms of sample variances, there is an equivalent formulation that is simpler for computation. Let $a_i(j)$ denote the rank of the $j$th sample with respect to attribute $i$. The equivalent method is to select the sample for which $n_{1[a_1(j)]} + n_{2[a_2(j)]}$ is minimised. This result follows from the fact that minimising $V$ is equivalent to minimising the sum of squares of the $n_{i[j]}$, because their sum is fixed, and the increase in sum of squares as a result of selecting sample $j$ is

$$\sum_{i=1}^{2}\left\{\left(n_{i[a_i(j)]}+1\right)^2 - \left(n_{i[a_i(j)]}\right)^2\right\}.$$

For the example above, we have

| Leaf number, $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Upper surface rank, $a_1[j]$ | 2 | 1 | 3 | 4 | 5 |
| Lower surface rank, $a_2[j]$ | 5 | 1 | 3 | 2 | 4 |
| $n_{1[a_1(j)]} + n_{2[a_2(j)]}$ | 9 | 10 | 9 | 8 | 10 |

again indicating that leaf 4 should be selected.

## 4 Simulations

It is straightforward to use simulation to compare the different methods of selecting samples. Aspects of the simulation that could be varied include the number of attributes and their population distribution, the set size, and the number of cycles. The impact of various types of ranking error could also be investigated. However, our intention here is to report the results of simulations under just a few conditions, to illustrate the performance of the methods, rather than to undertake a wide ranging study.

To this end, we fix the set size at $m = 5$ and the number of cycles at $r = 6$ (the spray deposit study had $m = 5$ and $r = 5$, but to implement method 2 we need an even number of cycles). The underlying population distribution is assumed to be multivariate normal.

Table 1 shows summary statistics from 200 simulations of bivariate normal variables with mean zero, variance one and correlation of either $\rho = 0$ or $\rho = 0.8$. Ranking is assumed to be perfect. The theoretical relative precision for a set size of $m = 5$ is 2.770 (Dell and Clutter, 1972). Each simulation yielded the number of samples $r_j$ at each rank for each variable, where $\sum_{j=1}^{m} r_j = 30 (= n)$. The relative precision for the given values of $r_j$ was then calculated as

$$RP = \frac{1}{n}\sigma^2 \Bigg/ \frac{1}{m^2}\sum_{j=1}^{m}\frac{\sigma_j^2}{r_j} ,$$

where $\sigma^2$ is the population variance (here equal to one) and $\sigma_j^2$ is the variance of the $j$th order statistic from a random sample of size $m$ from the population distribution (Dell and Clutter, 1972). In this calculation, the variances were fixed at their true values, but the values of the $\{r_j\}$ varied from simulation to simulation.

Table 1 shows the mean, minimum and maximum, and upper and lower quartiles ($Q_{25}$ and $Q_{75}$) of the $RP$ values for methods 1, 2 and 4. The values for method 1 apply only to the variable whose ranks are *not* used for selecting samples; the ranked set sample is balanced with respect to the other variable, and the $RP$ value for this variable is therefore 2.77. The values for methods 2 and 4 apply to both variables. The maximum values of $RP$ are sometimes slightly greater than 2.77, because balanced allocation is *not* the optimal allocation for a normally distributed response variable. Therefore an unbalanced allocation will sometimes arise by chance that gives a higher $RP$ value than the balanced allocation. In this example, the optimal (Neyman) allocation gives $RP = 2.797$.

For the values of $m$ and $r$ used in the simulations, the factor $c$ in equation 2 is 0.869, implying that when $\rho = 0$, the theoretical $RP$ values for methods 1 and 2 are 2.41 and 2.59 respectively; the mean values from the simulations are close to these values. However, method 4 performs much better, with a mean $RP$ value of 2.75 and a lower quartile of 2.73. Thus with this method of allocation, there is little loss of precision for either variable.

*(Table 1 about here)*

When $\rho = 0.8$ there is some improvement in the distribution of $RP$ values for methods 1 and 2, particularly in terms of the lower quartile. This is because, when the variables are correlated, the allocation for the variable that is not used in the ranking tends to be better than random. However, the improvements remain modest unless the correlation is close to one.

Suppose now that $K$ attributes are of interest, where $K \geq 2$. Methods 1–4 all extend directly, though for method 2 we require that the number of cycles is a multiple of $K$. The performance of methods 1 and 3 is unchanged and the lower bound for the relative precision for method 2 decreases to $[1 + (K-1)c]/K$. For method 4, the relative precision tends to decline as $K$ increases, because it is more difficult to obtain samples that are well balanced with respect to all $K$ variables. However, the decline is surprisingly slow, as shown in the second column of Table 2, which gives the mean $RP$ values from 200 simulations for $K = 2, \ldots, 6$ based on uncorrelated standard normal random variables.

*(Table 2 about here)*

Another aspect of method 4 that we have not yet discussed is the number of samples, $S_0$ say, that are required to ensure that there is at least one sample at each rank for each of the $K$ variables. For $K = 2$, it can be shown that $S_0 \leq 2m$ when $m \leq 6$, as will usually be the case in practice. However, for larger values of $m$ this bound does not hold; for example, for $m = 8$, $S_0$ can be as large as 17. It appears difficult to obtain general results, particularly for $K > 2$, but Table 2 gives some empirical results, summarising the mean, standard deviation and maximum value of $S_0$ from the 200 simulations. As is expected, these values increase with $K$, but again the increases are not rapid, and even with $K = 6$, $S_0$ never exceeded 15, and was less than 10 on average. Thus, in practice, method 4 is almost certain to include samples at all ranks for all variables, unless $n$ is very small.

## 5 Discussion

All of the methods that we have described are substantially better than using one variable as a concomitant for the other(s). When $\rho = 0$ this is equivalent to random sampling, so $RP = 1$, and when $\rho = 0.8$, $RP = 1.69$ when the population distribution is normal (from equation 3 of Ridout and Cobby, 1987, for example). In general, using one variable as a concomitant for another retains high efficiency only if there is a very high correlation between the variables. Thus, in most circumstances, it is well worth the additional effort of ranking each of the attributes of interest within each set.

As we have mentioned earlier, for very skewed distributions, an unbalanced allocation may be preferable to a balanced allocation. However, amongst the methods discussed, method 4 is clearly the best when a balanced allocation is desirable and when all attributes are of equal interest. Indeed, the loss of precision for this method is so slight that the method is appropriate even if one attribute is of somewhat greater interest than the others.

Even with the computational simplification of method 4 that was described in Section 3, it is easiest to use a computer program to do the calculations. Portable computers make this feasible, even if ranking is done in the field. Although there is some small extra labour in entering ranks into the computer, information about the ranks of the selected samples would normally be entered into a computer at some stage anyway. Entering the complete set of ranks for all attributes also allows measures of rank correlation between the different attributes to be calculated.

## Acknowledgement

This paper is based on an invited talk that I gave at a SPRUCE Advanced Workshop on Environmental Sampling and Monitoring held in Estoril, Portugal on 22-24 March 2001.

I am grateful to the organisers for inviting me to the Workshop. I am also grateful to two anonymous referees for their comments and suggestions.

## References

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

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

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

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

Norris, R.C.; Patil, G.P. and Sinha, A.K. (1995) Estimation of multiple characteristics by ranked set sampling methods. *Coenoses*, **10**, 95–111.

Patil, G.P.; Sinha, A.K. and Taillie, C. (1994a) Ranked set sampling. In: G.P. Patil and C.R. Rao (eds.), *Handbook of Statistics, Vol. 12: Environmental Statistics*, North Holland Elsevier, New York, 167–200.

Patil, G.P.; Sinha, A.K. and Taillie, C. (1994b) Ranked set sampling for multiple characteristics. *International Journal of Ecology and Environmental Sciences*, **20**, 94–109.

Ridout, M.S. and Cobby, J.M. (1987) Ranked set sampling with non-random selection of sets and errors in ranking. *Applied Statistics*, **36**, 145–152.

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.

Takahasi, K. (1970) Practical note on estimation of population means based on samples stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, **22**, 421–428.

**Table 1**. Summary statistics for relative precision ($RP$), from 200 runs of ranked set sampling using different methods of sample selection. The theoretical $RP$ for balanced ranked set sampling is 2.770 and for the Neyman allocation is 2.797.

| Method | Min | $Q_{25}$ | Mean | $Q_{75}$ | Max |
|---|---|---|---|---|---|
| | | | $\rho = 0$ | | |
| 1 | 1.28 | 2.27 | 2.42 | 2.65 | 2.78 |
| 2 | 2.20 | 2.50 | 2.58 | 2.69 | 2.78 |
| 4 | 2.70 | 2.73 | 2.75 | 2.77 | 2.78 |
| | | | $\rho = 0.8$ | | |
| 1 | 1.29 | 2.43 | 2.51 | 2.67 | 2.78 |
| 2 | 2.27 | 2.56 | 2.62 | 2.70 | 2.78 |
| 4 | 2.63 | 2.74 | 2.75 | 2.77 | 2.78 |

**Table 2**. Summary statistics for relative precision ($RP$) and for the number of samples needed to obtain at least one sample of each variable at each rank ($S_0$), from 200 runs of ranked set sampling with $K$ variables, using method 4 to select samples.

| $K$ | Mean $RP$ | Mean $S_0$ | SD $S_0$ | Max $S_0$ |
|---|---|---|---|---|
| 2 | 2.754 | 6.1 | 0.60 | 9 |
| 3 | 2.746 | 7.1 | 1.00 | 10 |
| 4 | 2.738 | 7.8 | 1.10 | 12 |
| 5 | 2.733 | 8.7 | 1.35 | 13 |
| 6 | 2.728 | 9.6 | 1.48 | 15 |

## BIOGRAPHICAL SKETCH

Martin Ridout is Reader in Statistics at the University of Kent, Canterbury. He obtained an MSc in Statistics from the University of Reading and then worked for many years as a statistical consultant in agricultural and horticultural research institutes. His current interests are in the analysis of discrete data and in the application of statistics to problems in biology, including genetics and yeast prion research.