Modelling ranking errors in ranked set sampling using a generalisation of Mallows' ϕ-model

Martin Ridout
Horticulture Research International - East Malling,
WEST MALLING, Kent ME19 6BJ, U.K.

***Present address**: Institute of Mathematics and Statistics, University of Kent at Canterbury, CANTERBURY, Kent, CT2 7NF, U.K.*

SUMMARY

Mallows' ϕ-model (Mallows, 1957, *Biometrika*, **44**, 114-130) is a simple one-parameter distribution for ranking data. Various generalisations of the model have appeared in the literature. Here we describe a different generalisation that is a mixture model. The new distribution is shown to be more successful than the basic ϕ-model in describing visual rankings of spray deposits on leaves.

1. Introduction

Suppose that a set of m objects is to be ranked independently by a population of judges on the basis of an attribute that can also be measured objectively. The objective measurement gives rise to a true ranking of the objects. We shall assume that there are no ties in the true ranking, and that ties are not allowed in the rankings produced by the judges. Mallows (1957) proposed several probability distributions to describe the rankings of the judges in relation to the true ranking. The simplest of these models, termed the ϕ-model, defines a distance measure between the observed ranking and the true ranking, and assumes that the probability of the observed ranking is a function of this distance. The distance measure is related to Kendall's τ (Kendall, 1970). Chung & Marden (1991) have pointed out that Mallows' ϕ-model appears earlier, albeit peripherally, in the work of Mann (1945), in accordance with Stigler's Law of Eponymy (Stigler, 1980).

Feigin & Cohen (1978) and Schulman (1979) investigated the model further, giving tables to facilitate parameter estimation, and Diaconis (1988) suggested the use of

alternative distance measures. Fligner & Verducci (1986) investigated a different type of generalisation of the ϕ-model, in which ranking is assumed to proceed in m-1 independent stages. At each stage, the object thought to rank highest amongst those remaining is removed. A separate parameter is associated with each stage, and the probability that the object selected at that stage is truly the highest ranked is a function of this parameter. When all m-1 parameters are equal, the model reduces to Mallows' ϕ-model.

In this paper we consider a different generalisation of Mallows' ϕ-model, motivated by a practical example of ranking errors in ranked set sampling. The model is a two-parameter mixture distribution. The paper is organised as follows. In Section 2 we describe Mallows' ϕ-model. Section 3 describes ranking data that arose during an evaluation of ranked set sampling for estimating spray deposits on the leaves of apple trees. We show that Mallows' model is not appropriate for some of these data. In Section 4 we generalise Mallows' model and show that the generalised model provides a better fit to the leaf deposit rankings. Relationships with other models of ranking errors are discussed briefly in Section 5.

2. Mallow' ϕ-model

We assume that the objects are labelled in accordance with their true ranking, so that the first object has the lowest rank, and the m^{th} object has the highest rank. Let $\pi(i)$ be the rank assigned to the i^{th} object, i=1, ..., m. Because ties are not allowed, $\pi = (\pi(1), ..., \pi(m))$ is a permutation of 1, ..., m. We shall define a distance function, $D(\pi)$, to be the minimum number of adjacent transpositions required to transform π to the identity permutation. A minimal sequence of adjacent transpositions may be generated by moving each object in turn to its correct place (this is the basis of the generalisation of Mallows' model due to Fligner & Verducci, 1986). For example, if m=5 and π=(2,3,1,5,4) a minimal sequence of adjacent transpositions is

$$(3,2,1,5,4) \rightarrow (3,1,2,5,4) \rightarrow (1,3,2,5,4) \rightarrow (1,2,3,5,4) \rightarrow (1,2,3,4,5)$$

and $D(\pi)$=4. For m objects, the possible values of $D(\pi)$ are 0, 1, ..., $m(m-1)/2$. Kendall's rank correlation between π and the identity permutation is a simple function of $D(\pi)$

(Kendall, 1970, pp. 7-8). We shall limit our attention to this particular distance measure, though, as pointed out earlier, alternative measures can be used.

In Mallows' ϕ-model, all permutations π with the same value of $D(\pi)$ are assumed to occur with equal probability

$$P(\pi) = \frac{\phi^{D(\pi)}}{\psi(\phi)} \tag{1}$$

where $\psi(\phi)$ is a normalising factor which ensures that the probabilities of all $m!$ permutations sum to one (Mallows used a different, but equivalent parameterisation). The parameter ϕ may take any non-negative real value. Perfect ranking, for which π is the identity permutation with probability one, corresponds to the limit $\phi \to 0$. When ϕ=1, all permutations are equally likely, and the ranking is random. When ϕ>1 the probability of a particular ranking *increases* with increasing $D(\pi)$. Although this is possible in principle, it seems unlikely to occur in practice, at least in the type of example that we have in mind, and we therefore assume that ϕ lies in the interval [0,1]; Feigin & Cohen (1978) impose the same constraint.

Since all permutations with the same value of D are assumed to be equally likely, our main interest is in the distribution of D itself. Clearly

$$P(D = d) = \frac{n_d \phi^d}{\psi(\phi)}, \qquad d = 0,1,...,m(m-1)/2 \tag{2}$$

where n_d is the number of permutations for which $D=d$. The computation of n_d is described by Kendall (1970, pp. 67-68). The mean and variance of D are

$$E(D) = \frac{m\phi}{1-\phi} - \sum_{j=1}^{m} \frac{j\phi^j}{(1-\phi^j)}$$

$$\text{var}(D) = \frac{m\phi}{(1-\phi)^2} - \sum_{j=1}^{m} \frac{j^2\phi^j}{(1-\phi^j)^2}$$

(Fligner & Verducci, 1986). Given a sample of D values, the maximum likelihood estimator of ϕ, $\hat{\phi}$, is obtained by solving the equation $E(D) = \overline{D}$, where $\overline{D}$ is the sample mean; Feigin & Cohen (1978) and Schulman (1979) give tables to facilitate estimation.

3. Ranking errors in ranked set sampling

Ranked set sampling is a technique for improving precision in comparison with simple random sampling, particularly with regard to estimation of the population mean. Kaur *et al*. (1995) provide an introduction to the literature on ranked set sampling. The method involves selecting *mr* random samples of size *m* from the population. Each set of *m* units is ranked in some way that does not involve direct measurement, for example by visual assessment, and one unit is measured from each set. This is done in such a way that, amongst the *mr* units that are measured in total, there are *r* that were thought to be of rank *j*, for each $j = 1, ..., m$.

Our interest here is in errors that may occur in the ranking process. As part of a recent practical evaluation of ranked set sampling for estimating spray deposits on the leaves of apple trees (Murray, Ridout & Cross, 2000), two plots of trees were sprayed with a fluorescent marker dye. One plot was sprayed at high volume, using coarse nozzles on the sprayer, to give a large average droplet size, the other was sprayed at low volume, using fine nozzles, to give a small average droplet size. We refer to these treatments as Coarse and Fine respectively. One hundred and twenty five leaves were sampled haphazardly from each set of trees. In the laboratory the 125 leaves were divided randomly into 25 sets of five. Four observers, working independently, then ranked each of these sets by assessing visually the density of spray deposit. The upper and lower leaf surfaces were ranked separately. The 125 leaves were then divided randomly again into 25 sets of five and the procedure was repeated. This gave in total 800 rankings of m=5 leaves (25 sets × 2 sprayer settings × 2 surfaces × 4 observers × 2 repetitions). The true ranking of each set is also known, based on an objective measurement of deposit density using an image analysis system.

For the purposes of analysis we pooled the data from the two repetitions to give 16 sets of rankings (2 sprayer settings × 2 surfaces × 4 observers), with 50 rankings in each set. We fitted Mallows' ϕ-model to each set. The mean value of D ranged from 0.98 to 1.50, implying a range of 0.220 to 0.321 for $\hat{\phi}$. However, there was some systematic lack of fit. Table 1 shows the ratio of the sample variance to the variance of the fitted Mallows' distribution for the 16 data sets. For all observers the ratio was less than one for ranking

the upper surfaces of leaves from the Fine treatment. Otherwise the ratio was consistently greater than one, sometimes substantially so. In the next section we develop a mixture model that can account for this overdispersion.

4. A generalisation of Mallows' model

In ranked set sampling, our interest is in the performance of a particular ranker, who has to rank several different sets of samples. For a particular set, the parameter ϕ determines the probabilities of the different possible rankings. However, the difficulty of ranking is likely to vary from set to set, and this would be reflected in variation of ϕ from set to set. We therefore consider a mixture distribution in which ϕ is replaced by a random variable *T*. Since ϕ is constrained to lie in the interval [0,1], it is flexible and convenient to assume a beta distribution for *T*. Alternatives, such as the logistic-normal distribution, in which logit(ϕ) is assumed to be normally distributed, could also be considered. However, the resulting mixture distribution of *D* is determined primarily by the low-order moments of the mixing distribution, and is likely to be relatively insensitive to the precise distributional form of the mixing distribution.

We therefore assume that the probability density function of *T* is

$$f(t) = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 \le t \le 1, \qquad \alpha, \beta > 0$$

where $B(\alpha,\beta)$ denotes the beta function, $\alpha = \phi(1-\delta)/\delta$, $\beta = (1-\phi)(1-\delta)/\delta$ and δ is an overdispersion parameter $(0 \le \delta \le 1)$. This implies that

$$E(T) = \phi \qquad \text{and} \qquad \text{var}(T) = \delta\phi(1-\phi)$$

The resulting distribution of *D* is

$$P(D=d) = \frac{n_d}{B(\alpha,\beta)} \int_0^1 \frac{t^{d+\alpha-1}(1-t)^{\beta-1}}{\psi(t)} dt, \qquad d = 0,1,...,m(m-1)/2 \qquad (3)$$

Often, in mixture models that involve the beta distribution, reasonably simple explicit expressions exist for the mixture distribution probabilities. Examples are the beta-binomial distribution (*e.g.* Morgan, 1992, Section 6.1.1) and the beta-geometric distribution (*e.g.* Morgan, 1992, Section 6.7.2). Here, however, explicit expressions for

the probabilities appear not to exist, due to the presence of the normalising factor $\psi(t)$ in the denominator of the integrand. Similarly, explicit formulae for the moments do not exist. Nonetheless, it is easy to work with the distribution numerically, particularly when *m* is small, as it usually is in ranked set sampling.

The generalised distribution was fitted to the 12 data sets for which the variance ratio in Table 1 exceeded one. In six instances there was a significant improvement in fit ($P<0.05$) when judged by a likelihood ratio test. Table 2 shows the observed and fitted values for the rankings by Observer 1 of the upper and lower leaf surfaces from the Coarse treatment. Parameter estimates and likelihood ratio statistics are shown in Table 3. The calculations were done using Genstat. The standard errors are based on a numerical approximation to the observed information matrix. For the ϕ-model the standard error of $\hat{\phi}$ calculated in this way agreed very closely with the value calculated from a formula based on the expected information (Fligner & Verducci, 1986).

Judged by the likelihood ratio test, the improvement in fit is strongly significant for the upper leaf surface ($P=0.003$) but not for the lower leaf surface ($P=0.11$). Nonetheless, the two data sets show a similar pattern of discrepancies from the ϕ-model and the lack of significance for the lower leaf surface data probably reflects the lack of power of the test statistic, with a sample size of 50, rather than the adequacy of the ϕ-model. For both data sets, the Wald statistic that results from dividing $\hat{\delta}$ by its standard error indicates a less significant departure from the ϕ-model than the likelihood ratio statistic does. This was true for all 12 data sets studied and probably results in part from skewness in the distribution of $\hat{\delta}$. The estimates of ϕ were similar for both models, but, as would be expected, the standard error was greater for the mixture model.

We have focused our attention so far on the distribution of *D*. If we wish to consider the distribution of individual permutations we need to assess the validity of the assumption that all permutations with the same value of *D* are equiprobable. This is difficult to do directly because, for most values of *D*, the number of rankings is smaller than the number of distinct permutations, and we therefore adopt an indirect approach. In ranked set sampling the most important information about ranking errors is contained in the matrix

$P = (p_{ij})$ where p_{ij} is the probability that the object that has true rank i within a set is judged to have rank j. From the observed data we may calculate the matrix $N = (n_{ij})$ where n_{ij} is the number of times that the object that has true rank i within a set was judged to have rank j. Table 4 shows the matrix N for the rankings by Observer 1 of the upper leaf surface of leaves from the Coarse treatment. Each of the rows and columns of this matrix sums to 50, the total number of rankings. Table 4 also shows the fitted values $\hat{n}_{ij} = 50\hat{p}_{ij}$, based on the fit of the new distribution for D (4th column of Table 2) and assuming that all rankings with the same value of D are equiprobable. The matrix of fitted values is necessarily symmetric about both diagonals.

The chi-squared statistic for comparing the observed and fitted values is X^2=14.3. This statistic has 14 degrees of freedom (d.f.), because the row and column totals of N are fixed, and because two parameters have been estimated. The fit is therefore satisfactory in this important respect. If the standard Mallows' model (3rd column of Table 2) is used instead, the goodness of fit statistic is X^2=24.3 (15 d.f.). For the rankings of the lower leaf surface (last two columns of Table 2) the corresponding statistics are X^2=16.1 (14 d.f.) for the new model and X^2=23.0 (15 d.f.) for the ϕ-model.

5. Discussion

Mallows' ϕ-model is extremely ambitious, in so far as it attempts to describe the distribution of m! rankings in terms of a single parameter. It is not surprising, therefore, that examination of the fit of the model to particular data sets has often revealed inadequacies of the model (*e.g.* Cohen & Mallows, 1983). The mixture model described here introduces a second parameter to allow for overdispersion. This gives considerable additional flexibility in modelling the distribution of D. In some instances modelling D may be an end in itself but, if the probabilities of individual rankings are required, the new model still has the restrictive assumption that all rankings with the same value of D are equiprobable. However, in the particular example that we have studied, the new model leads to good estimates of the matrix P, which is derived from individual permutation probabilities.

The new model assumes that variability in the difficulty of ranking different sets leads to variability in ϕ. In some circumstances it may be useful to regard the differences between sets as fixed effects. For example, ϕ could be related to a measure of within-set variability via a logistic model. In the context of ranked set sampling, however, it is more appropriate to regard set differences as a random effect, with the sets considered as a random sample from the population of possible sets that might need to be ranked.

Several other models of ranking errors have been considered in the literature of ranked set sampling. Patil *et al*. (1994) used a model proposed originally by Luce (1959) and Bohn & Wolfe (1994) modelled the matrix *P* directly. However, the model of Dell & Clutter (1972) has received the most attention. This model assumes that an observer ranks the units on the basis of his or her *perception* of the true value. The perceived value is assumed to be the sum of the true value and a random error term that is distributed as $N(0,\sigma^2)$. This is a form of the Thurstone-Mosteller-Daniels (TMD) ranking model (Daniels, 1950). Under this model, rankings with the same value of *D* are *not* equiprobable in general. Nonetheless, we may still consider the distribution of *D* under this model. Limited simulations suggest that this distribution is *underdispersed* relative to Mallows' ϕ-model. This is in agreement with a numerical example in Feigin & Cohen (1978). These authors simulated a sample of 40 rankings using a TMD model and noted that the ϕ-model gave a poor fit. The ratio of the observed variance to the variance of the ϕ-model with the same mean was 0.45 in their example.

The fact that, for all observers, the variance ratio for ranking the deposits on the upper surfaces of leaves from the Fine treatment was less than one suggests that a TMD model may be appropriate for describing ranking errors. Conversely, the overdispersion present in the other data sets indicates that a TMD model would *not* be appropriate for these data.

ACKNOWLEDGEMENT

I am grateful to Prof. Byron Morgan for helping to identify a numerical error in this work.

REFERENCES

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

Chung, L. & Marden, J.I. (1991). Use of nonnull models for rank statistics in bivariate, two-sample, and analysis of variance problems. *Journal of American Statistical Association*, **86**, 188-200.

Cohen, A. & Mallows, C.L. (1983). Assessing goodness of fit of ranking models. *The Statistician*, **32**, 361-373.

Daniels, H.E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society B*, **12**, 171-181.

Dell, T.R. & Clutter, J.L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, **28**, 545-553.

Diaconis, P. (1988). *Group Representations in Statistics*. Institute of Mathematical Statistics Lecture Notes-Monograph Series, No. 11.

Feigin, P.D. & Cohen, A. (1978). On a model of concordance between judges. *Journal of the Royal Statistical Society B*, **40**, 203-213.

Fligner, M.A. & Verducci, J.S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society B*, **48**, 359-369.

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

Kendall, M.G. (1970). *Rank Correlation Methods*, 4th edn. London: Charles Griffin & Co.

Luce, R.D. (1959). *Individual Choice Behaviour*. New York: John Wiley.

Mallows, C.L. (1957). Non-null ranking models. *Biometrika*, **44**, 114-130.

Mann, H.B. (1945). Nonparametric tests against trend. *Econometrica*, **13**, 245-259.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. London: Chapman & Hall.

Murray R.A., Ridout, M.S. & Cross J.V. (2000) The use of ranked set sampling in spray deposit assessment. *Aspects of Applied Biology*, **57**, 141-146.

Patil, G.P., Sinha, A.K. & Taillie, C. (1994). Ranked set sampling for multiple characteristics. *International Journal of Ecology & Environmental Sciences*, **20**, 357-373.

Schulman, R.S. (1979). Ordinal data: an alternative distribution. *Psychometrika*, **44**, 3-20.

Stigler, S. (1980) Stigler's Law of Eponymy. *Transactions of the New York Academy of Sciences. Series 2*, **39**, 147-157.

Table 1

The ratio of the sample variance to the variance of the fitted Mallows' ϕ-model for 16 sets of ranking data.

Spray Treatment	Surface	Observer 1	2	3	4
Coarse	Upper	1.75	1.12	1.61	1.63
Coarse	Lower	1.33	1.53	1.41	1.27
Fine	Upper	0.86	0.71	0.95	0.85
Fine	Lower	1.30	1.30	1.23	1.73

Table 2

Observed distribution of D *and fitted values from the standard Mallows' ϕ-model and the generalised distribution described in Section 4 of the paper. The rankings are by Observer 1 and relate to leaves from the Coarse treatment.*

D	Upper leaf surface			Lower leaf surface		
		Fitted			Fitted	
	Observed	ϕ-model	New model	Observed	ϕ-model	New model
0	22	16.5	22.4	19	15.9	19.2
1	14	17.2	12.4	16	17.0	14.5
2	6	10.0	7.0	6	10.2	8.3
3	4	4.3	3.9	4	4.6	4.3
4	0	1.5	2.2	4	1.6	2.1
5	3	0.4	1.2	1	0.5	1.0
6-10	1	0.1	0.9	0	0.2	0.6

Table 3

Parameter estimates for the data sets in Table 2, and likelihood ratio (LR) statistics for comparing models. Figures in brackets are asymptotic standard errors.

	Upper leaf surface	Lower leaf surface
ϕ-model		
$\hat{\phi}$	0.260 (0.0320)	0.267 (0.0324)
New model		
$\hat{\phi}$	0.252 (0.0411)	0.262 (0.0377)
$\hat{\delta}$	0.190 (0.0832)	0.103 (0.0788)
LR statistic (1 d.f.)	8.78	2.52

Table 4

The rankings of Observer 1 in relation to the true rankings for the 50 sets of upper leaf surfaces from the Coarse treatment. The main entries are the observed frequencies. The figures in brackets are the frequencies predicted from fitting the new model to the observed distribution of D*, and assuming that all rankings with the same value of* D *are equiprobable.*

	Observer's ranking				
True rank	1	2	3	4	5
1	43	4	1	1	1
	(37.9)	(7.8)	(2.6)	(1.1)	(0.6)
2	6	30	9	3	2
	(7.8)	(31.5)	(7.1)	(2.5)	(1.1)
3	0	11	30	8	1
	(2.6)	(7.1)	(30.6)	(7.1)	(2.6)
4	1	4	6	32	7
	(1.1)	(2.5)	(7.1)	(31.5)	(7.8)
5	0	1	4	6	39
	(0.6)	(1.1)	(2.6)	(7.8)	(37.9)

Notes on data organisation

32 sets of ranking data generated, in file **:\USER\RANKSS\1998DATA\PERMS.DAT**. Each data set therefore has 25 values.

Data sets	Observer
1-8	1
9-16	2
17-24	3
25-32	4

Within each set of 8 the order is as follows:

Data set	Treatment	Surface	Replicate
1	Coarse	Upper	1
2	Coarse	Upper	2
3	Coarse	Lower	1
4	Coarse	Lower	2
5	Fine	Upper	1
6	Fine	Upper	2
7	Fine	Lower	1
8	Fine	Lower	2

Within the 32 data sets, each of the 25 lines of data is a permutation of (1,2,3,4,5). A typical permutation, such as (1,4,2,5,3) means the following (**NB - rank 1 = lowest, rank 5 = highest, this is different from the paper, which uses the reverse convention**).

The sample thought to be lowest was indeed lowest (rank 1).

The sample thought to be 4^{th} lowest was actually 2^{nd} lowest

The sample thought to be 2^{nd} lowest was actually 3^{rd} lowest

The sample thought to be 5^{th} lowest was actually 4^{th} lowest

The sample thought to be 3^{rd} lowest was actually 5^{th} lowest

The Genstat program is as follows:

```
unit [125]


fact [lev=25] Set1C, Set2C, Set1F, Set2F
fact [lev=5] Rank1CU[1...4], Rank1CL[1...4], \
             Rank1FU[1...4], Rank1FL[1...4], \
             Rank2CU[1...4], Rank2CL[1...4], \
             Rank2FU[1...4], Rank2FL[1...4], \
             True1CU, True1CL, True1FU, True1FL, \
             True2CU, True2CL, True2FU, True2FL

open 'o:/user/rankss/1998Data/rankss4.dat'; ch=2
read [ch=2; pr=*] SampleNo, Set1C, Rank1CU[1...4], Rank1CL[1...4], \
            C1Upper, C1Lower, True1CU, True1CL
read [ch=2; pr=*] SampleNo, Set2C, Rank2CU[1...4], Rank2CL[1...4], \
            C2Upper, C2Lower, True2CU, True2CL
read [ch=2; pr=*] SampleNo, Set1F, Rank1FU[1...4], Rank1FL[1...4], \
            F1Upper, F1Lower, True1FU, True1FL
read [ch=2; pr=*] SampleNo, Set2F, Rank2FU[1...4], Rank2FL[1...4], \
            F2Upper, F2Lower, True2FU, True2FL
close 2


for x = (True1CU, True2CU, True1CL, True2CL, True1FU, True2FU, \
        True1FL, True2FL)4; \
    y = Rank1CU[1], Rank2CU[1], Rank1CL[1], Rank2CL[1], \
        Rank1FU[1], Rank2FU[1], Rank1FL[1], Rank2FL[1], \
        Rank1CU[2], Rank2CU[2], Rank1CL[2], Rank2CL[2], \
        Rank1FU[2], Rank2FU[2], Rank1FL[2], Rank2FL[2], \
        Rank1CU[3], Rank2CU[3], Rank1CL[3], Rank2CL[3], \
        Rank1FU[3], Rank2FU[3], Rank1FL[3], Rank2FL[3], \
        Rank1CU[4], Rank2CU[4], Rank1CL[4], Rank2CL[4], \
        Rank1FU[4], Rank2FU[4], Rank1FL[4], Rank2FL[4]; \
    s = ( (Set1C, Set2C)2, (Set1F, Set2F)2)4
    "tabu [class=x,y; print=nobs] SampleNo"
    calc z = y
    tabu [class=s,x] z; mean=m
    print [rlprint=*; clprint=*; ipr=*; sq=y] m; d=0; f=1
    print [sq=y] ':'
endfor

stop
```