Hacettepe Journal of Mathematics and Statistics
Volume 40 (6) (2011), 907 – 913

REFINING THE JUDGES' ASSESSMENTS OF ITEMS IN PAIRED COMPARISON EXPERIMENTS

Nasir Abbas[*†], Muhammad Aslam[‡] and Zawar Hussain[‡]

Received 24 : 08 : 2010 : Accepted 20 : 06 : 2011

Abstract

In the method of paired comparisons, items are compared on the basis of their qualitative characteristics assessed by judges through their sensory evaluations. Judges are offered items in pairs and are asked to pick the better one. The experiment is repeatedly executed to yield preference data based on binary digits – zeros and ones – allotted to the items by the judges. The preferred item is awarded rank one while the loser is assigned zero. As the binary digits fail to furnish the actual comparative worth of items and indistinguishably assign one to the preferred item and zero to the losing one, a methodology is proposed to measure the actual comparative worth of the competing items on a finer scale by assigning some refined rank on a finer scale to each of the two competing items. The assigned ranks are then converted to a refined paired comparison data-set in the form of a preference matrix to be used for ranking the items. For illustration, a real data-set on ice-cream brands is used to rank the brands using the renowned Bradley-Terry model for paired comparisons.

Keywords: Paired Comparisons, Bradley-Terry Model, Refined data-set, Maximum likelihood (ML) estimates, Iterative procedure, Test of Goodness of Fit.

2000 AMS Classification: 62 J 15, 62 F 10.

[*] Department of Statistics, Government Postgraduate College, Jhang, Pakistan.
E-mail: nabbasgcj@yahoo.com

[†] Corresponding Author.

[‡] Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.
E-mail: (M. Aslam) aslamsdqu@yahoo.com (Z. Hussain) zhlangah@yahoo.com

1. Introduction

In the method of paired comparisons (PC), judges (raters, respondents, jurists, panelists etc.) are presented with items (stimuli, options, objects, items, individuals etc.) in pairs and are asked to choose the better one on the basis of sensory evaluations according to certain criteria. The experiment is repeatedly executed to yield preference data which is presented in a preference matrix and is used via the paired comparison models to rank items, which is the ultimate goal of the method of paired comparisons. The method is widely employed for comparing items when no real measurement of worth (strength, merit, etc.) of items is possible. The method is utilized in a variety of applications ranging from sensory testing to the investigations of preferences, sports and choice behavior, etc.

Customarily, the paired comparison experiment is repeatedly executed to yield preference data based on binary digits – zeros and ones – allotted to items by judges. The preferred item is awarded rank one while the loser is given zero. As the binary digits – zero and one – fail to furnish the actual comparative worth of items and indistinguishably assign one to the preferred item and zero to the loser. In this article, a methodology is proposed to measure the actual comparative worth of the competing items by assigning a specific rank on a finer scale to each of the competing items. The assigned ranks are then converted to the usual paired comparison data-set in the form of a preference matrix to be used for ranking the items. Though all the PC models may serve the purpose, we stick only to the renowned Bradley-Terry (BT) model proposed by Bradley and Terry [3].

The literature reveals different situations in which the method of PC is used, and also discusses various models devoted to ranking items in these situations. For instance, David [6] provides a detailed review of the PC models. Bradley [4] assumes the responses follow the Logistic distribution and proposes his model, whereas Abbas and Aslam [1] adopt the Cauchy distribution. The motivation for using refined ranks in the method of paired comparisons is to accommodate the actual worth of the items under study. An improvement in the method of paired comparisons has been proposed by Abbas and Aslam [2] through the accommodation of quantitative weights in qualitative paired comparisons.

We may break up this study as follows: Section 2 discusses the theory for the proposed methodology of the refined paired comparisons. Section 3 deals with the iterative ML estimation procedure and Section 4 provides an illustrative example using a real dataset on five brands of ice-cream coded as A, B, C, D and E. Section 5 pertains to the comparison of the conventional and the refined paired comparisons. Section 6 concludes the entire study.

2. The refined paired comparison technique

As we know that we usually assign one-zero ranks to the items under study in paired comparison experiments. But it does not properly accommodate the degree of liking or disliking of the judges. If, in the paired comparison of items-pair (i, j), a judge prefers item i to j by assigning it rank one, and similarly in the paired comparison of the items-pair (i, k), item i is preferred to k by being assigned rank one, then there is no criterion for differentiating between the worth of item i in comparison to item j and that of i with k. It necessitates items j and k being alike in their worth and that i is preferable to both j and k with same degree of liking. But actually the item i may enjoy different (weaker or stronger) preferences over items j and k. But the conventional paired comparison plan merely prefers one item to the other and fails to perceive degree of liking or disliking of judges. It, therefore, necessitates a refinement in the judges' assessments of the items under study to properly accommodate the so-called degree of superiority or inferiority of one item over the other. This is exactly what we attempt to articulate in this research. Such a refinement of the paired comparison experiments may also be incorporated if judges are asked to make more than one paired comparison for the pair (i, j), so that the actual strength of the items may be incorporated properly.

We know that a confusing situation may emerge while comparing three items A, B and C pair-wise. A judge may prefer item A to B, B to C and then C to A. This is technically known as a circular-triad that exhibits inconsistency of the judges. Such a situation may arise when a judge is either incompetent in assessing the difference between items or the items are too similar to be distinguished. But in the refined paired comparison technique, the occurrences of circular-triads may be avoided to a considerable extent. Here the judges will have to assign ranks to items on a finer scale to locate the actual worth of items on the basis of their liking or disliking, and the chance of occurrence of circular-triads will be minimized. However, if the judges are competent enough to asses the differences occurring among the items under study, the chance of the occurrence of circular-triads may be completely ruled out.

Numerically, the range of values of the response variable is enhanced from the binary values $(0, 1)$ to $0, 1, 2, \ldots, m$. Here the value '0' stands for the zero or minimum strength or worth of an item and m denotes the maximum strength. Consequently, while comparing items i and j, a judge may assign any value ranging from 0 to m depending upon its strength or worth as perceived by the judge. These values are used while comparing all $C_2^t = t(t-1)/2$ possible pairs if there are, in all, t items under study. The refined data thus obtained are them converted back to the conventional paired comparison data by omitting a pre-specified percentage of the middle values if ties are to be accommodated. These values may generate a paired comparison data-set with ties, which is studied by Rao and Kupper [8], Davidson [7] and many others. We may use the judges' assessments falling within a certain range $[L, U]$ as ties. The set of value falling below L may be assumed as being against a specific item and those greater than U may be regarded as favorable to the other one.

3. The ML estimation

For illustration, we make use of the renowned BT model for paired comparisons which states:

$$\phi_{ij} = H(V_i - V_j) = \tfrac{1}{4}\int_{-(V_i - V_j)}^{\infty} \sec h^2\left(\tfrac{1}{2}x\right) dx = \frac{\pi_i}{\pi_i + \pi_j}, \tag{3.1}$$

for all $i\ (\neq j) = 1, \ldots, t$. Here t denotes the total number of items to be compared, $V_i = \log(\pi_i)$, $\pi_i > 0$, denotes the worth or merit of the item i for $i = 1, \ldots, t$, ϕ_{ij} represents the probability of preferring item i over j and $\phi_{ji} = 1 - \phi_{ij}$, David [6].

Here we note that the outcomes fall into just two possible categories, i.e., either i will be preferred to j or vice versa, the experiment of paired comparisons between any two items i, j is independently performed for a fixed number of times n_{ij} (say), where $n_{ij} = n_{ji}$ is the number of comparisons made between items i and j for $i\ (< j) = 1, 2, \ldots, t$, and $a_{ij} + a_{ji} = n_{ij} = a'_{ij} + a'_{ji}$, were a_{ij} and a'_{ij} respectively symbolize the conventional and the refined number of preferences of item i over j. Hence the variable $\boldsymbol{A}$ denoting the number of preferences of item i over j follows a binomial distribution, i.e.,

$$P\left(\boldsymbol{A} \,|\, \boldsymbol{\pi}\right) = \frac{n_{ij}}{a_{ij}!a_{ji}!}\left(\phi_{ij}\right)^{a_{ij}}\left(\phi_{ji}\right)^{a_{ji}} = \frac{n}{a_{ij}!a_{ji}!}\left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{a_{ij}}\left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{a_{ji}},$$

where $a_{ji} = n_{ij} - a_{ij}$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_t)$, $\boldsymbol{A} = \{a_{ij}\}$, $\phi_{ji} = 1 - \phi_{ij}$.

Using the classical *ML* estimation method to estimate the parameters of the Bradley-Terry model, the likelihood function $L(\boldsymbol{A}; \boldsymbol{\pi})$ is given by

$$L(\boldsymbol{A}; \boldsymbol{\pi}) = \prod_{i(<j)=1}^{t} \left[\frac{n}{a_{ij}!a_{ji}!} \left(\frac{\pi_i}{\pi_i + \pi_j} \right)^{a_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j} \right)^{a_{ji}} \right]$$

$$\Longrightarrow L(\boldsymbol{A}; \boldsymbol{\pi}) \propto \prod_{i=1}^{t} \left[(\pi_i)^{a_i} \right] \prod_{i(<j)=1}^{t} \left[(\pi_i + \pi_j)^{-n_{ij}} \right], \tag{3.2}$$

where $a_i = \sum_{j(\neq i)=1}^{t} a_{ij}$ denotes the total score of item i for $i = 1, 2, \ldots, t$. Since, the logarithmic function is non-decreasing and both the likelihood function and its logarithmic form are maximized at the same points (estimates); after dropping the terms independent of parameters, the logarithmic form of (3.2) is:

$$\ln L(\boldsymbol{A}; \boldsymbol{\pi}) = \sum_{i=1}^{t} \{a_i \ln(\pi_i)\} - \sum_{i<j}^{t} \{n_{ij} \log(\pi_i + \pi_j)\}. \tag{3.3}$$

Following the algebraic maximization technique, we equate to zero the partial derivatives of (3.3) with regards to the unknowns π_i and get

$$\frac{a_i}{\hat{\pi}_i} - \sum_{j \neq i}^{t} \left(\frac{n_{ij}}{\hat{\pi}_i + \hat{\pi}_j} \right) = 0. \tag{3.4}$$

Equations (3.4) are complicated and an analytic solution for the worth parameters is intractable. So, following Bradley [4, 5], we plan to find the *ML* estimates through iterative methods.

If $\pi_i^{(k)}$ be the k-th approximation of π_i, then $\pi_i^{(k)} = \pi_i^{*(k)} \Big/ \sum_{i=1}^{t} \pi_i^{*(k)}$, where

$$\pi_i^{*(k)} = a_i \left\{ \sum_{j \neq i=1}^{t} \left(\frac{n_{ij}}{\pi_i^{(k-1)} + \pi_j^{(k-1)}} \right) \right\}^{-1}.$$

Here, for identification we impose the restriction that the sums of the worth parameters be unity. For the initialization of the iterative procedure, we may take $\pi_i^{(0)} = 1/t$, for all $i = 1, 2, \ldots, t$, and repeat the iterations till convergence. The ML estimates $\hat{\pi}_i$ of the worth parameters π_i may be used to rank the set of t items.

4. Numerical illustration

We consider an example of a real data-set on five brands of ice-cream coded as A, B, C, D and E. The brands were offered in pairs to a set of 20 students of a local college. The students were asked to assign refined ranks as a result of sensory evaluations to the competing brands on a linear scale having range 1 through 5 on the basis of the characteristics the brands possess. The students did the job in the stipulated pattern. The refined ranks a'_{ij} assigned to the items-pair (i, j) by all the judges, summed over j to yield the total scores of all the brands of ice-cream under consideration, along with the conventional data matrix are summarized in Table 1.

Table 1. Conventional and refined data-sets

Ice-cream brands	A	B	C	D	E	a_i, (a'_i)
A	0	16 (83)	13 (95)	15 (92)	12 (89)	56 (359)
B	4 (37)	0	6 (27)	11 (81)	8 (39)	29 (184)
C	7 (32)	14 (73)	0	7 (25)	9 (42)	37 (172)
D	5 (23)	9 (56)	13 (57)	0	7 (33)	34 (169)
E	8 (33)	12 (69)	11 (63)	13 (85)	0	44 (250)

The numbers within parentheses indicate the refined observations obtained by summing the ranks a'_{ij} having individual values in the pre-set range 1 to 5 for all the 20 students, while those without parentheses correspond to the conventional ones. The last column denotes the conventional and the refined scores, a_i and a'_i, of the ice-cream brands under consideration. We may solve (3.4) using the usual iterative procedure in the light of the data given in Table 1. But, here we run a computer program developed in *SAS* package using the procedure *PROC GENMOD*. The *SAS* codes are given in the Appendix. The resulting estimates, along with the associated standard errors (SEs) within parentheses, are displayed in Table 2. The difference occurring in the worth of the competing brands by switching the conventional paired comparisons over to the refined ones are summarized in the last column.

Table 2. Estimates and SEs using conventional and refined data-sets

Ice-cream brands	π_i	π'_i	Differences in worth
A	0.5227 (0.2993)	0.6729 (0.1267)	0.1502
B	-0.6307 (0.2953)	-0.5552 (0.1244)	0.0755
C	-0.2915 (0.2897)	-0.4203 (0.1270)	0.1288
D	-0.4171 (0.2911)	-0.6224 (0.1247)	0.2053
E	0.0000 (0.0000)	0.0000 (0.0000)	0.0000

From the conventional estimates of worth (column 2), it is evident that the set of ice-cream brands under study may be ranked as A being the number one, E the second, C the third, D the fourth and B the fifth and last one. However, in the light of the model's estimates obtained using the refined data-set, all the brands enjoy the same ranking order except that of the brands B and D, which have their ranks reversed, i.e., B captures the fourth position and D the last and fifth one. Moreover, the brands A and B are over estimated by 0.1502 and 0.0755 respectively via the refined ranks, C and D are in order under-estimated by 0.1288 and 0.2053. However, the worth of the brand E remains intact for both the data-sets. We may get the overall worth of any brand by taking the geometric means of its worth estimates found using the conventional and the refined data-sets as $\sqrt{\pi_i\pi'_i}$. It is interesting to note that we obtain smaller standard errors for the estimates of the worth parameters using the refined data-set.

5. Comparison of the conventional and the refined paired comparison techniques

We witness that the ranking order is preserved for the ice-cream brands A, E and C for the paired comparison data-sets collected through both of the conventional as well as the refined assessments, but the ranking order for the brands B and D is reversed for the data-sets collected through the refined assessments. Moreover, we obtain smaller standard errors for the estimates of the worth parameters using the refined data-set, which indicates an improvement in the paired comparison technique.

6. Concluding remarks

The Bradley-Terry model is the most popular model in paired comparison studies and is extensively used for ranking items. Because of its simple mathematical form, it can be easily handled for analytical estimation of the worth parameters. It occupies a vital position in this field. Since we have obtained the refined data-set by paying more attention to using more careful assessments, we get smaller standard errors. So, it may be said that the refined technique may preferably be used in place of the conventional one for a number of reasons. Firstly, it captures a more accurate assessment of the worth of items being compared than does the conventional assessment in the form of mere zeros and ones. It utilizes some additional information regarding the worth of the items under study, which is desirable in any statistical analysis. Secondly, the smaller standard errors observed for the refined data-set as compared to those for the conventional data-set indicate a move in the right direction towards an improvement in the paired comparison technique. Thirdly, if the paired comparison data-sets obtained using the refined and the conventional assessments of the items under consideration coincide, then the two approaches become identical.

Appendix

```
data creams; input win n A B C D E;

datalines;
16 20 1 -1 0 0 0
13 20 1 0 -1 0 0
15 20 1 0 0 -1 0
12 20 1 0 0 0 -1
6 20 0 1 -1 0 0
11 20 0 1 0 -1 0
8 20 0 1 0 0 -1
7 20 0 0 1 -1 0
9 20 0 0 1 0 -1
7 20 0 0 0 1 -1
;
proc genmod;
model win/n = A B C D E / dist=bin link=logit noint;
run;
```

References

[1] Abbas N. and Aslam, M. *Prioritizing the items through paired comparison models, a Bayesian approach*, Pakistan Journal of Statistics **25**(1), 59–69, 2009.

[2] Abbas, N. and Aslam, M. *Extending the Bradley-Terry model for paired comparisons to accommodate weights*, Journal of Applied Statistics **38** (3), 571–580, 2011.

[3] Bradley R. A. and Terry, M. E. *Rank analysis of incomplete block designs: I. The method of paired comparisons*, Biometrika **39**, 324–345, 1952.

[4] Bradley, R. A. *Some statistical methods in taste testing and quality evaluation*, Biometrics **9** (1), 22–38, 1953.

[5] Bradley, R. A. *Paired comparisons: Some basic procedures and examples*, In: Krishnaiah, P. R., Sen, P. K. (Eds.) (Handbook of Statistics **4**, North-Holland, Amsterdam, 1984), 299–326.

[6] David, H. A. The Method of Paired Comparisons, 2nd Ed (Charles Griffin & Company Ltd., London, 1988).

[7] Davidson, R. R. *On extending the Bradley–Terry model to accommodate ties in paired comparison experiments*, Journal of American Statistical Association **65**, 317–328, 1970.

[8] Rao, P. V. and Kupper, L. L. *Ties in paired comparison experiments: A generalization of the Bradley-Terry model*, Journal of the American Statistical Association **62** (317), 194–204, 1967.