Significance Testing in Bradley Terry Models

Rohit Patel

Marketing Sciences Group
IMRB International, Delhi

Contents

Note: Readers who are already aware of the underlying model/techniques and/or need only to see the part that deals with implementation and usage of the program can skip Sections 2-4 and can directly proceed to Section 5.

1 Introduction

One often has to address questions involving comparison among products. Experiments are designed based on various considerations of practical relevance and analytical feasibility. We focus on analysis of the data from experiments that deal with respondents' preference between products. A typical response is the selection of a preferred product between two given products. This is like cumulative results of a sequence of round-robin tournaments involving the products. We assume homogeneity of respondents (theoretical considerations suggest that proper design of experiment can greatly reduce the effects of non-homogeneity of respondents). Our intent is to understand the products so as to be able to answer questions like "Which is better?", "What is the margin of difference?" and "Which of them are equally preferred?". For this we assume that these preferences can be characterized mathematically as 'product abilities', as we call them. Hence a product with a higher ability (or strength) is more likely to be preferred over a product with a lower ability in a comparison between the two. These abilities are realized and then hypotheses tests are carried out. A typical hypothesis to be tested is that of the equality of all abilities. This is equivalent to saying "All products are equally preferred". Other more complex hypotheses can be tested that say "Abilities of this group of products are equal". We use Likelihood Ratio Tests on Bradley-Terry model specification to test the hypotheses. This has a distinct advantage over conventional testing methods as the entire data is used for any hypothesis that is being tested. The program is with IMRB International, Delhi. You may write to us for implementation which can be done reasonably quickly.

2 The Model

Consider an experiment with 't products' involving 'paired comparisons'. We begin with the following assumptions to specify the model (Bradley and Terry, 1952)

- The products have true ratings, or abilities. We denote these by π_i for the i^th product. Furthermore, let $\pi_i \geq 0 \;\; \forall i$ and $\Sigma\pi_i = 1$ without any loss of generality.

- In a contest between product i and product j, the probability that i is preferred over j is $\pi_i/(\pi_i + \pi_j)$

- Every contest between any two products is independent of others, and that they are under identical conditions.

The model can alternatively be expressed in logit–linear form

$$\text{logit}[\ \text{pr}(i \text{ beats } j)] = \lambda_i - \lambda_j$$

The values λ_i, λ_j, etc., occur on a linear scale and permit over–all comparisons of the product abilities. Any consideration of the differences among product abilities should be based on the values of $\log(\pi_i)'s$.

3 The Likelihood Function

We can now write down the likelihood function for a given outcome as follows, let n_{ij} denote the number of times product i is preferred over product j.

$$L = \prod_i \left(\frac{\pi_i}{\pi_i + \pi_j} \right)^{n_{ij}}$$

In the case where r_{ijk} denotes the rank of i^{th} product in k^{th} repetition, repetitions being equal for all product pairs (say n), the likelihood function can be written as,

$$L = \prod_i \pi_i^{2n(t-1) - \sum_{i \neq j} \sum_k r_{ijk}} \prod_{i<j} (\pi_i + \pi_j)^{-n}.$$

4 Estimation and Likelihood Ratio Tests

Consider the hypothesis of the following kind,

$$H_0 : \pi_i = 1/t \qquad (i = 1, ..., t)$$

$$H_1 : \pi_i = \pi(h) \qquad (h = 1, ..., m);$$

$$i = s_{h-1} + 1, ..., s_h, \text{ where } s_0 = 0, s_m = t \text{ and } \sum_h (s_h - s_{h-1})\pi(h) = 1,$$

If $p(h)$ is the maximum likelihood estimate of $\pi(h)$, these estimates are obtained from the equations

$$\left[\left\{2n(t-1)(s_h - s_{h-1}) - \sum_{i=sh-1+1}^{s_h} \sum_{j\neq i} \sum_k r_{ijk} - \frac{1}{2}n(s_h - s_{h-1})(s_h - s_{h-1} - 1)\right\} \Big/ p(h)\right]$$

$$-n(s_h - s_{h-1}) \sum_{f\neq h} (s_f - s_{f-1})/\{p(h) + p(f)\} = 0$$

and $\sum_h (s_h - s_{h-1})p(h) = 1$

The likelihood ratio (LR)can now be written down, and the likelihood ratio test can be performed. If the number of repetitions are sufficiently large, $-2\log(LR)$ follows a χ^2 distribution with $(m-1)$ degrees of freedom.

We will now shift our focus to a wider class of hypotheses, of the following kind,

$$H_0 : \pi_i = \pi(h) \quad (h = 1, ..., m), \quad \pi(h_1) = \pi(h_2) = ... = \pi(h_k);$$

$$i = s_{h-1} + 1, ..., s_h$$

against

$$H_1 : \pi_i = \pi(h) \qquad (h = 1, ..., m);$$

$$i = s_{h-1} + 1, ..., s_h,$$

where $s_0 = 0, s_m = t$ and $\sum_h (s_h - s_{h-1})\pi(h) = 1$

How the test is performed in this case is that we first find the MLE of the $\pi(h)$'s under the null and the alternate hypothesis and then we calculate the $-2\log(LR)$ for the hypothesis. Under the assumption of sufficient number of repetitions this follows a χ^2 distribution with $(k-1)$ degrees of freedom. The program calculates the value of the statistic and then gives the χ^2 values at 95% and 99%, it also gives the P-value for the hypothesis.

5 Implementation, a general hypothesis

Suppose you have 10 products ($P_1, P_2, \ldots, P_{10}$) at hand, with some predetermined grouping. Grouping here means that all the products in a group have same abilities. There can be many reasons you have such grouping. For instance if you already know that products in each group are same, or if you have done other tests to conclude that these groups exist. In an experiment involving soft drinks, one can have reasons to believe that all cola CSD's will have similar abilities or all pineapple based juices will have similar abilities. Lastly, such grouping can be formed with the use of this program itself as you will see.

$$\overbrace{P_1, P_2, P_3}^{Group1} \overbrace{P_4, P_5}^{Group2} \overbrace{P_6, P_7}^{Group3} \overbrace{P_8, P_9}^{Group4} \overbrace{P_{10}}^{Group5}$$

And you want to test if Group 2 and Group 5 have same abilities. So, under the null hypothesis ($H0$) you will put them in the same group and under the alternate hypothesis ($H1$) you will do the grouping as you originally believe.

$$H0 : \overbrace{P_1, P_2, P_3}^{Group1} \overbrace{P_4, P_5, P_{10}}^{Group2} \overbrace{P_6, P_7}^{Group3} \overbrace{P_8, P_9}^{Group4}$$

$$H1 : \overbrace{P_1, P_2, P_3}^{Group1} \overbrace{P_4, P_5}^{Group2} \overbrace{P_6, P_7}^{Group3} \overbrace{P_8, P_9}^{Group4} \overbrace{P_{10}}^{Group5}$$

You will be able to test this kind of $H0$ against $H1$.
Note: Any k groups can be clubbed to form a group in $H0$, in above example, $H0$ can be of the following form.

$$H0 : \overbrace{P_1, P_2, P_3}^{Group1} \overbrace{P_4, P_5, P_{10}, P_6, P_7}^{Group2} \overbrace{P_8, P_9}^{Group3}$$

6 Data Input

We first need to create an input data file, the program takes data input in csv format. One way to do this is to open 'Microsoft Excel' and input the win-loose count as shown in Figure 1. The first two columns is the products between which the comparison was made and the third column gives the number of times the 'winner' was preferred over the 'loser'. There are $t = 4$ products in the example, generating a total of $2! \times \binom{t}{2} = 12$ rows. Finally Save As → CSV(Comma delimited)(*.csv). The first column should have title "winner", the second one "loser" and the third one "Freq" with a capital F, all are case sensitive. Leave blank the rows for which comparisons are not available, or you can just delete these rows, it makes no difference.

	A	B	C
1	winner	loser	Freq
2	MB31	K732	15
3	MB31	MU59	14
4	MB31	D419	17
5	K732	MB31	35
6	K732	MU59	18
7	K732	D419	20
8	MU59	MB31	36
9	MU59	K732	32
10	MU59	D419	20
11	D419	MB31	33
12	D419	K732	30
13	D419	MU59	30
14			

Figure 1: Input of the Win-Loose Count

	A	B
1	MB31	a
2	K732	b
3	MU59	b
4	D419	d
5		

2(a)

	A	B
1	MB31	a
2	K732	b
3	MU59	c
4	D419	d
5		

2(b)

Figure 2: An example, Entering the null(2(a)) and the alternate(2(b))

7 Null and alternate hypothesis

Once you have entered data, it's time to enter null and alternate hypothesis. Again, the program accepts csv files. Figure 2 shows a certain $H0$ which says products "K732" and "MU59" have same ability, against, $H1$ that all 4 products have different abilities. Figure 3 shows how the null hypothesis in the example in Section 5 will be entered.

8 Running the program

8.1 Quick Run

You can use the "quick_run.cmd" script to quickly run the program. For this:
– Name the Win-Loose count file "wlc.csv"

	A	B
1	P1	Group1
2	P2	Group1
3	P3	Group1
4	P4	Group2
5	P5	Group2
6	P6	Group3
7	P7	Group3
8	P8	Group4
9	P9	Group4
10	P10	Group2
11		

	A	B
1	P1	Group1
2	P2	Group1
3	P3	Group1
4	P4	Group2
5	P5	Group2
6	P6	Group3
7	P7	Group3
8	P8	Group4
9	P9	Group4
10	P10	Group5
11		

Figure 3: Entering the hypotheses in Section 5

– Name the null hypothesis file "h0.csv"
– Name the alternate hypothesis file "h1.csv"

Now 'double-click' the file "quick_run.cmd" and the output will be saved in the file "Output.txt".

8.2 The Run Script

You can also run the program using the "run.cmd" script, suppose your Win-Loose Count filename is 'file1.csv' and your $H0$ and $H1$ file names are 'file2.csv' and 'file3.csv' respectively. Then you first run the "run.cmd" script, this will open a command window with R terminal. You need to run the following commands in the terminal

```
> source("BTSig")
> BTSig("file1.csv","file2.csv","file3.csv")
```

9 Few Examples

We now look at some typical tests that one might want to do

9.1 Are all abilities equal?

In this case, under the null hypothesis all products will be in the same group. And in the alternate hypothesis they all will be in different group. Suppose

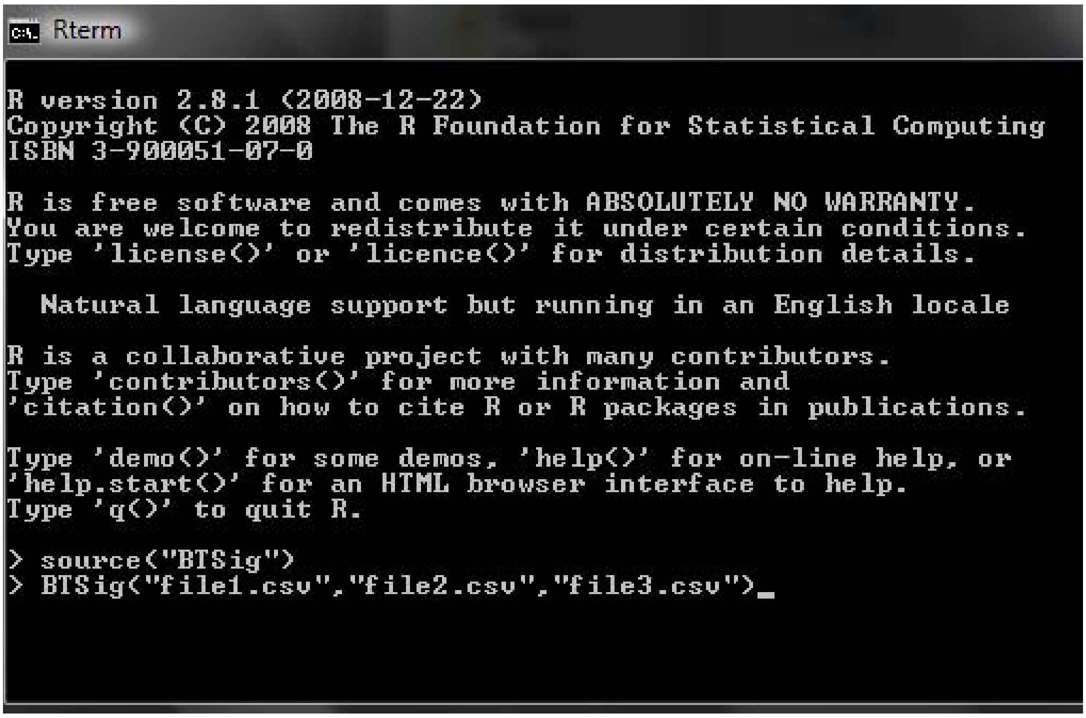

Figure 4: Using the "run.cmd" script

there are t products.

$$H0 : \overbrace{P_1, P_2, P_3, \ldots, P_t}^{Group1}$$

$$H1 : \overbrace{P_1}^{Group1} \quad \overbrace{P_2}^{Group2} \quad \overbrace{P_3}^{Group3} \; \ldots\ldots \; \overbrace{P_{t-1}}^{Group(t-1)} \quad \overbrace{P_t}^{Group(t)}$$

9.2 Is the best same as the rest?

Here, our null hypothesis is that all abilities are equal, but we are testing it against $H1$ that the null is different from the rest, assuming that the rest of them are the same. Suppose the abilities show that P_h is the best, then

$$H0 : \overbrace{P_1, P_2, P_3, \ldots, P_t}^{Group1}$$

$$H1 : \overbrace{P_h}^{Group1} \quad \overbrace{P_1, P_2, \ldots, P_{h-1}, P_{h+1}, \ldots, P_t}^{Group2}$$

9.3 Is the best same as the second best?

Here, the null hypothesis is that the best is the same as second best, and we put no constraints on the rest of them. Suppose P_h is the best and P_k is the second best as shown by the abilities, then

$$H0: \overbrace{P_h, P_k}^{Group(h)} \quad \overbrace{P_1}^{Group1} \ \ldots \ \overbrace{P_{h-1}}^{Group(h-1)} \ \overbrace{P_{h+1}}^{Group(h+1)} \ \ldots \ \overbrace{P_{k-1}}^{Group(k-1)} \ \overbrace{P_{k+1}}^{Group(k+1)} \ \ldots \ \overbrace{P_t}^{Group(t)}$$

$$H1: \overbrace{P_1}^{Group1} \quad \overbrace{P_2}^{Group2} \quad \overbrace{P_3}^{Group3} \ \ldots\ldots \ \overbrace{P_{t-1}}^{Group(t-1)} \quad \overbrace{P_t}^{Group(t)}$$

9.4 Is this grouping valid?

You can check if a particular grouping makes sense by checking it against the alternate of inequality, like in the example in Section 7.

$$H0: \overbrace{P_1, P_2 \ldots P_{s_1}}^{Group1} \overbrace{P_{s_1+1}, \ldots, P_{s_2}}^{Group2} \ldots \overbrace{P_{s_{m-1}+1}, \ldots, P_{s_m}}^{Group(m)}$$

$$H1: \overbrace{P_1}^{Group1} \quad \overbrace{P_2}^{Group2} \quad \overbrace{P_3}^{Group3} \ \ldots\ldots \ \overbrace{P_{t-1}}^{Group(t-1)} \quad \overbrace{P_t}^{Group(t)}$$

10 Understanding the Output

The output of the example in Sections 6 and 7 (ref Figure 1 and 2) is as follows:

```
$Abilities
      D419       K732       MB31       MU59
 0.0000000 -0.4185212 -0.9965700 -0.1060473

$pi_estimates
     D419      K732      MB31      MU59
0.3416999 0.2248451 0.1261363 0.3073188

$Abilities_under_alternate
      D419       K732       MB31       MU59
 0.0000000 -0.4185212 -0.9965700 -0.1060473

$pi_estimates_under_alternate
     D419      K732      MB31      MU59
```

```
0.3416999 0.2248451 0.1261363 0.3073188

$Abilities_under_null
      D419       K732       MB31       MU59
 0.0000000 -0.2619686 -0.9934038 -0.2619686

$pi_estimates_under_null
     D419      K732      MB31      MU59
0.3437153 0.2645010 0.1272826 0.2645010

$Chi_Square_Statistic_Value
[1] 2.337454

$Chi_Square_at_99
[1] 6.634897

$Chi_Square_at_95
[1] 3.841459

$P_Value
[1] 0.1262958

$Verdict
[1] "We do not reject the null hypothesis at 95%"
```

10.1 $Abilities

This gives the Abilities of all the products assuming that all products are distinct with different abilities, this is useful to determine which is the best product, or what grouping can you test for. These are λ_i values. We set $\lambda_1 = 0$ for convenience of comparison, hence, $\lambda_i = \log \pi_i - \log \pi_1$.

10.2 $pi_estimates

These are the π_i values assuming the products are all distinct.

10.3 $Abilities_under_alternate

These are the ability estimates assuming the groupings under alternate. You will see all products in a group have same abilities. *(In our example, the* *alternate specifies a distinct group for every product and hence different abilities for all products)*

10.4 $pi_estimates_under_alternate

These are the π_i estimates assuming the groupings under alternate. You will see all products in a group have same estimates. *(In our example, the alternate specifies a distinct group for every product and hence different estimates for all products)*

10.5 $Abilities_under_null

These are the ability estimates assuming the groupings under null. You will see all products in a group have same abilities. *(In our example, the null specifies same group for K732 and MU59)*

10.6 $pi_estimates_under_null

These are the π_i estimates assuming the groupings under null. You will see all products in a group have same estimates.

10.7 $Chi_Square_Statistic_Value

This is the final calculated statistic value, if this exceeds the cutoff point, we reject the hypothesis.

10.8 $Chi_Square_at_99

The cutoff at $\alpha = 99\%$ is being given for convenience. This is the χ^2 value at 99%.

10.9 $Chi_Square_at_95

The cutoff at $\alpha = 95\%$.

10.10 $P_Value

This is the P Value of the test, or the significance.

10.11 $Verdict

Self explanatory.

The above example shows us that D419 is much more preferred over the rest, and MB31 is the least. The hypothesis tested here is the one in the example in Section 7, Figure 2. Also, we see from the example that there is not enough evidence to reject the hypothesis that "Products "K732" and "MU59" have same ability".

For questions, suggestions and implementation please contact

Rohit Patel	Rohit.Patel@imrbint.com
Sudhanshu Gaur	Sudhanshu.Gaur@imrbint.com