PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

REMARKS ON THE METHOD OF PAIRED COMPARISONS: III. A TEST OF SIGNIFICANCE FOR PAIRED COMPARISONS WHEN EQUAL STANDARD DEVIATIONS AND EQUAL CORRELATIONS ARE ASSUMED*

FREDERICK MOSTELLER
HARVARD UNIVERSITY

A test of goodness of fit is developed for Thurstone's method of paired comparisons, Case V. The test involves the computation of

$$\chi^2 = n\Sigma(\theta'' - \theta')^2/821,$$

where n is the number of observations per pair, and θ'' and θ' are the angles obtained by applying the inverse sine transformation to the fitted and the observed proportions respectively. The number of degrees of freedom is $(k-1)(k-2)/2$.

1. *Introduction*

It would be useful in Thurstone's method of paired comparisons to have a measure of the goodness of fit of the estimated proportions to the observed proportions. Ideally we might try to find estimates of the stimuli positions S_i such that we can reproduce the observed proportions p'_{ij} as closely as possible in some sense.

One kind of test might be based on

$$\chi^2 = \frac{\Sigma(p''_{ij} - p'_{ij})^2}{\sigma^2_{ij}}$$

where p''_{ij} is the estimate of p'_{ij} derived from the S'_i. But the true p_{ij} are not known and would have to be replaced by the observed p'_{ij}. If one does replace the p_{ij} by p'_{ij} and σ_{ij} by σ'_{ij}, then it is possible to fit the S'_i by means of a minimum chi-square criterion. However, such a procedure calls for an iterative scheme and involves extremely tedious computations. An alternative method is suggested by the inverse sine transformation.

*This research was performed in the Laboratory of Social Relations under a grant made available to Harvard University by the RAND Corporation under the Department of the Air Force, Project RAND.

2. *The model*

It is assumed that we have a set of stimuli which, when presented to a subject, produce sensations. These sensations are assumed to be normally distributed, perhaps with different means. However the standard deviations of each distribution are assumed to be the same, and the correlations between pairs of stimuli sensations are assumed equal.

Subjects are presented with pairs of stimuli and asked to state which member of each pair is greater with respect to some property attributed to all the stimuli (the property is the dimension of the scale we are trying to form). Our observations consist of the proportions of times stimulus j is judged "greater than" stimulus i. We call these proportions p'_{ij} to indicate that they are observations and not the true proportions p_{ij}.

From the observed proportions we compute normal deviates X'_{ij} and proceed in the usual way (5) to estimate the stimulus positions, S'_i, on the sensation scale. Once the S'_i are found we can retrace to get the fitted normal deviates X''_{ij} and the fitted proportions p''_{ij}.

Our problem is to provide a method for ascertaining how well the fitted p''_{ij} agree with the observed p'_{ij}.

In such a test of significance involving goodness of fit, we are interested in knowing what the null hypothesis and the alternative hypothesis are. In the present case the null hypothesis is given by the model assumed above. However, the alternative hypothesis is quite general: merely that the null hypothesis is not correct. In particular, the null hypothesis assumes additivity so that if D_{ij} is the distance from S_i to S_j and D_{jk} is the distance from S_j to S_k, we should find

$$D_{ik} = D_{ij} + D_{jk}\,.$$

If we do not have unidimensionality this additivity property will usually not hold.

For example, consider the case of three stimuli with $S_1 < S_2 < S_3$. If the standard deviation of each distribution is the same, we might write

$$\begin{aligned} D_{12} &= S_2 - S_1 \\ D_{13} &= S_3 - S_1 \\ D_{23} &= S_3 - S_2\,. \end{aligned}$$

Since we can choose $S_1 = 0$ and $S_2 = D_{12}$, S_3 from the second equation must be D_{13}. Finally

$$D_{23} = D_{13} - D_{12}\,.$$

Since each of our comparisons of stimuli is done independently it is not necessary that this relation hold either for the observations or for the theoretical values. Indeed the observed value of D_{23} could have conflicted with the assumption of additivity. Such a failure of additivity makes the fitting of the observed p'_{ij} less likely, and on the average failure will increase the value of χ^2 in our test.

It can also happen that the standard deviations of the various stimuli are not equal even though unidimensionality obtains. In this case our attempt to fit the data under the equal standard deviations assumption will sometimes fail, and this failure will be reflected, in general, in a failure of additivity and thus an increase in χ^2.

3. *The transformation*

Like so many other good things in statistics, the inverse sine transformation was developed by R. A. Fisher (4). Further discussion by Bartlett (1, 2), Eisenhart (8), and Mosteller and Tukey (7) may be of interest to those who wish to examine the literature. The facts essential to the present discussion are these: If we have an observed p' arising from a binomial sample of size n from a population with true proportion of successes p, then

$$\theta' = \arcsin \sqrt{p'} \tag{1}$$

is approximately normally distributed with variance

$$\sigma_{\theta'}^2 = \frac{821}{n}, \tag{2}$$

nearly independent of the true p, when θ' is measured in degrees. A table for making the transformation to angles has been computed by C. I. Bliss (3), and is readily available in G. W. Snedecor's *Statistical Methods* (4th Edition), p. 450.

Then if we define

$$\begin{aligned} \theta'_{ij} &= \arcsin \sqrt{p'_{ij}} \\ \theta''_{ij} &= \arcsin \sqrt{p''_{ij}} \end{aligned} \tag{3}$$

where p'_{ij} are the observed proportions and p''_{ij} are the proportions derived from fitting the S_i, we can test goodness of fit by

$$\chi^2 = \sum_{i<j} \frac{(\theta''_{ij} - \theta'_{ij})^2}{821/n}. \tag{4}$$

If there are k stimuli we have k parameters to fit, the k S'_i values. But two of these are the zero point and the scale factor, which are arbitrary. This leaves k—2 parameters free for fitting the data. There are $k(k-1)/2\, p'_{ij}$'s to be fitted. So it appears that the appropriate number of degrees of freedom for the test is $k(k-1)/2 - (k-2) - 1 = (k-1)(k-2)/2$. We note that with two stimuli we can always fit the data perfectly, so there should be zero degrees of freedom as the formula indicates.

4. *Illustrative example*

To illustrate the test we will use the paired comparison method on the American League baseball record for 1948. The following table gives the observed p'_{ij}. The number in the ith row and jth column is the proportion of games won by the team named at the top of the jth column from the team named at the left of the ith row. In this situation we regard the clubs as stimuli which have distributions of performances. The number of games each club plays with each other club is 22 (except for minor fluctuations). Successive tables indicate the steps in the solution. The steps are these:

1. From p'_{ij} table obtain X'_{ij} table from a table of the normal integral.
2. Solve for the S'_i by summing columns and averaging.
3. Use S'_i to obtain X''_{ij}, $X''_{ij} = S'_i - S'_j$.
4. Use X''_{ij} to obtain p''_{ij}, from a table of the normal integral.
5. Compute θ'', θ', $\theta'' - \theta'$.
6. Get the sum of squares of $\theta'' - \theta'$.
7. Divide the sum of squares by $821/n$, here $821/22$.
8. Look up result in χ^2 table with $(k-1)(k-2)/2$ degrees of freedom.

PROPORTIONS OF ALL GAMES THAT THE TEAM GIVEN AT THE TOP OF THE COLUMN WON FROM THOSE AT THE LEFT (1948)

Each Entry Represents 22 Games

p'_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	—	.478	.545	.273	.409	.364	.273	.273
Bost.	.522	—	.364	.455	.318	.318	.318	.364
N.Y.	.455	.636	—	.455	.409	.273	.227	.273
Phil.	.727	.545	.545	—	.545	.182	.364	.273
Det.	.591	.682	.591	.455	—	.500	.273	.364
St.L.	.636	.682	.727	.818	.500	—	.545	.381
Wash.	.727	.682	.773	.636	.727	.455	—	.429
Chic.	.727	.636	.727	.727	.636	.619	.571	—

X'_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	——	—.055	+.113	—.604	—.230	—.348	—.604	—.604
Bost.	+.055	——	—.348	—.113	—.473	—.473	—.473	—.348
N.Y.	—.113	+.348	——	—.113	—.230	—.604	—.749	—.604
Phil.	+.604	+.113	+.113	——	+.113	—.908	—.348	—.604
Det.	+.230	+.473	+.230	—.113	——	.000	—.604	—.348
St.L.	+.348	+.473	+.604	+.908	.000	——	+.113	—.303
Wash.	+.604	+.473	+.749	+.348	+.604	—.113	——	—.179
Chic.	+.604	+.348	+.604	+.604	+.348	+.303	+.179	——
	+2.332	+2.173	+2.065	+0.917	+0.132	—2.143	—2.486	—2.990
S'_i	.2915	.2716	.2581	.1146	.0165	—.2678	—.3108	—.3738

X''_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	——							
Bost.	.0199	——						
N.Y.	.0334	.0135	——					
Phil.	.1769	.1570	.1435	——				
Det.	.2750	.2551	.2416	.0981	——			
St.L.	.5593	.5394	.5259	.3824	.2843	——		
Wash.	.6023	.5824	.5689	.4254	.3273	.0430	——	
Chic.	.6653	.6454	.6319	.4884	.3903	.1060	.0630	——

p''_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	——							
Bost.	.508	——						
N.Y.	.513	.505	——					
Phil.	.570	.562	.557	——				
Det.	.608	.601	.595	.539	——			
St.L.	.712	.705	.700	.649	.612	——		
Wash.	.726	.720	.715	.665	.628	.517	——	
Chic.	.747	.741	.736	.687	.652	.542	.525	——

Table of θ'', θ', $\theta'' - \theta'$

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.								
Bost.	45.46 46.26 —.80							
N.Y.	45.75 42.42 +3.33	45.29 52.89 —.760						
Phil.	49.02 58.50 —9.48	48.56 47.58 +0.98	48.27 47.58 +0.69					
Det.	51.24 50.24 +1.00	50.83 55.67 —4.84	50.48 50.24 +0.24	47.24 42.42 +4.82				
St.L.	57.54 52.89 +4.65	57.10 55.67 +1.43	56.79 58.50 —1.71	53.67 64.75 —11.08	51.47 45.00 +6.47			
Wash.	58.44 58.50 —.06	58.05 55.67 +2.38	57.73 61.55 —3.82	54.63 52.89 +1.74	52.42 58.50 —6.08	45.97 42.42 +3.55		
Chic.	59.80 58.50 +1.30	59.41 52.89 +6.52	59.08 58.50 +0.58	55.98 58.50 —2.52	53.85 52.89 +0.96	47.41 51.88 —4.47	46.43 49.08 —2.65	

$$\Sigma(\theta'' - \theta')^2 = 551.40$$

$$821/22 = 37.32$$

$$\chi^2_{21} = 14.78 \qquad .80 < P(\chi^2) < .90$$

The chi-square result shows rather good agreement between the fitted data and the observed data. Investigation of additional baseball data has suggested that the agreement is usually too good rather than not good enough. It was suggested to the author that a possible reason for this is that the proportion of games won by any team from another team involves an admixture of games played at home and away, and that if these were separated we might then not get such consistently good agreement. As an example, suppose probabilities of winning at home and away are .25 and .75 respectively, averaging .50. The variance of games won based on the $p = .50$ is $n/4$, but based on $n/2$ games at .25 and $n/2$ at .75, the variance is $3n/16$, somewhat smaller. The decrease in variance would be similar to that gained from stratified sampling. Calculations not presented here suggest that this may be the case.

It should be remembered that we have found the best S'_i's in the least-squares sense to reproduce the X'_{ij}'s, and have not done our best to reproduce the θ's. This means that, had we done a more elaborate method of fitting, we might have obtained a still better fit and consequently a higher value of P (which is already quite high).

5. *The power of the test for three stimuli*

The power of the test developed, that is the probability of rejecting the null hypothesis when it is false, is rather awkward to investigate. The power depends on the degree of divergence from the assumptions, the number of stimuli involved, the number of observations for each pair of stimuli, as well as the significance level chosen. We will discuss the power for a rather special case. This case has the advantage that it displays the workings of the chi-square test rather clearly and is easy to compute. Our procedure will be: (1) set up the model, (2) compute χ^2 for this case, (3) insert a departure from the model, (4) investigate the power for the special case under consideration.

We will assume that the standard deviations of the differences between pairs of stimuli are unity. The true stimuli means are in the order $S_3 > S_2 > S_1$. Furthermore we will assume that these means are sufficiently close to one another that the approximation

$$p_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-(S_j - S_i)}^{\infty} e^{-\frac{1}{2}x^2}\, dx \cong \frac{1}{2} + \frac{S_j - S_i}{\sqrt{2\pi}}\,, \tag{5}$$

will be adequate. For this case p_{ij} will be nearly 1/2, so we will be able to use the approximation:

$$\sigma^2(p'_{ij}) = \frac{1}{4n} = \sigma^2. \tag{6}$$

Working with this case will have the further advantage that we will not need to use the inverse sine transformation but can work directly with

$$\chi^2 = \frac{\sum\limits_{i<j} (p'_{ij} - p''_{ij})^2}{\sigma^2}, \tag{7}$$

since our principal reason for working with the transformation was that σ^2 was not known.

The observations can be written

$$p'_{ij} = p_{ij} + k_{ij}\sigma . \tag{8}$$

Here the unprimed p is the true proportion of the time stimulus j is reported to exceed stimulus i, the primed p is the corresponding observed proportion, σ is $1/4n$, and k_{ij} is a random normal deviate with zero mean and standard deviation unity. The sample size is n assumed to be reasonably large.

Under these assumptions

$$\begin{aligned} p'_{ij} = p_{ij} + k_{ij}\sigma &= \frac{1}{\sqrt{2\pi}} \int_{-(S_j - S_i)+\xi_{ij}}^{\infty} e^{-\frac{1}{2}x^2}\, dx \cong \frac{1}{2} + \frac{S_j - S_i - \xi_{ij}}{\sqrt{2\pi}} \\ &\cong \frac{1}{2} + \frac{S_j - S_i + k_{ij}\sigma\sqrt{2\pi}}{\sqrt{2\pi}} . \end{aligned} \tag{9}$$

Thus the normal deviate corresponding to p'_{ij} is approximately

$$D'_{ij} = S_j - S_i + k_{ij}\sigma\sqrt{2\pi} . \tag{10}$$

Now we insert these values in the paired comparison table as usual and solve for the estimates of the stimuli positions S'_i by summing columns and averaging. After adding the mean of the true stimuli positions these estimates are:

$$\begin{aligned} S'_1 &= S_1 - (k_{12} + k_{13})\sigma\sqrt{2\pi}/3 , \\ S'_2 &= S_2 + (k_{12} - k_{23})\sigma\sqrt{2\pi}/3 , \\ S'_3 &= S_3 + (k_{13} + k_{23})\sigma\sqrt{2\pi}/3 . \end{aligned} \tag{11}$$

We take the differences of these pairs to get the fitted normal deviates, the D''_{ij}:

$$\begin{aligned} D''_{12} &= S_2 - S_1 + (2k_{12} + k_{13} - k_{23})\sigma\sqrt{2\pi}/3 , \\ D''_{13} &= S_3 - S_1 + (k_{12} + 2k_{13} + k_{23})\sigma\sqrt{2\pi}/3 , \\ D''_{23} &= S_3 - S_2 + (-k_{12} + k_{13} + 2k_{23})\sigma\sqrt{2\pi}/3 . \end{aligned} \tag{12}$$

Now the fitted proportions p''_{ij} are approximately

$$p''_{ij} = \frac{1}{2} + \frac{D''_{ij}}{\sqrt{2\pi}}. \tag{13}$$

When we take the differences $p'_{ij} - p''_{ij}$ we get

$$\begin{aligned} p'_{12} - p''_{12} &= (k_{12} - k_{13} + k_{23})\sigma/3\,, \\ p'_{13} - p''_{13} &= (-k_{12} + k_{13} - k_{23})\sigma/3\,, \\ p'_{23} - p''_{23} &= (k_{12} - k_{13} + k_{23})\sigma/3\,. \end{aligned} \tag{14}$$

Now immediate computation of χ^2 inserting the values from equations (14) into equation (7) is

$$\chi^2 = \left(\frac{k_{12} - k_{13} + k_{23}}{\sqrt{3}} \right)^2 . \tag{15}$$

Since the k's are normally and independently distributed with zero means and unit variance, the quantity in parentheses is in turn a normal deviate with zero mean and unit variance, because the standard deviation of the sum in the numerator is $\sqrt{3}$. Of course, the square of such a normal deviate is distributed like χ^2 with one degree of freedom. In this special case then we have shown how the χ^2 test arises.

We have incidentally set up the machinery for examining the power of the test for our special case. Until now we have assumed that the p_{ij} were arranged to get consistency in the spacings between the true stimuli means. We now relax this condition. In particular let us suppose that the consistent p_{23} is replaced by $p_{23} + \Delta$ where Δ is an error due to the lack of unidimensionality of the stimuli we are considering. This means that p'_{23} will be replaced by $p'_{23} + \Delta$, which in turn means that k_{23} will be replaced by $k_{23} + \Delta/\sigma$. Now when we come to compute χ^2 with the null hypothesis not satisfied we get

$$\chi^{2*} = (\chi + \Delta/\sqrt{3}\,\sigma)^2. \tag{16}$$

Here χ is a normal deviate, the expression inside the parentheses on the right of equation (15). If we are working with a significance test at the 5% level we will reject the null hypothesis unless

$$-1.96 < \chi + \Delta/\sqrt{3}\,\sigma < 1.96\,.$$

The following table indicates very roughly how often we will reject the null hypothesis as $\Delta/\sqrt{3}\,\sigma$ takes various values.

$\Delta/\sqrt{3}\,\sigma$	Percent rejected
1	16%
1.96	50%
2	52%
3	84%

We say roughly because when Δ takes large values our approximations no longer hold very well. Nevertheless these values are indicative of the magnitudes.

Let us see how much error there must be in p_{23} to raise the rejection level to 16%. Suppose $\underline{n} = 48$. Then

$$\frac{\Delta^2}{3\sigma^2} = 1$$

$$\Delta^2 = 3\sigma^2 = \frac{3}{4 \times 48} = \frac{1}{64}$$

$$\Delta = .125 .$$

Thus for samples as large as 48, p_{23} must deviate from the consistent value of approximately .5 by as much as .125 to raise the probability of rejection from 5% to 16%.

A short discussion of the kinds of alternatives that can exist in paired comparisons and the general behavior of this test against these may assist the reader. The principal ways the Case V assumptions can be violated are

(1) lack of normality,

(2) lack of unidimensionality,

(3) failure of the equal standard deviation of differences assumption.

Failure of normality is not important to the method of paired comparisons, as we shall show elsewhere. It is just as well then that the present test will be very poor at detecting deviations from normality. The normality assumption is more in the nature of a computational device than anything else.

Lack of unidimensionality will be reflected in the failure of distances between estimated stimuli positions to agree with the observed distances, and thus we will have high chi-square values. The principal alternative of interest then, is one for which the test is sensitive.

Unfortunately it is also possible that we have unidimensionality without having equality of standard deviations of differences of pairs. The result of using Case V may be to give a large chi-square value when this happens. This is not uniformly true however. It is possible to have unequal standard deviations without detecting this fact in the Case V solution as has been shown elsewhere (6). In particular, if there is only one aberrant standard deviation, and if the stimulus mean for that stimulus is near the mean of all the stimulus positions, the chi-square test will not be likely to detect this failure of the model. The best that can be said is that sometimes such aberrations will cause high values of chi-square and sometimes not, depending on the nature of the case.

We might like to relax our conditions and not use Case V but try to use some other case. However, this requires a large number of stimuli. In the case of the assumption of independence between pairs of stimuli we still have for k stimuli a total of k means and k variances to choose. Two of these $2k$ values are merely scale and location parameters, so we have in all $2k - 2$ things that can be varied as against $k(k—1)/2$ cell entries. Thus we need at least 5 stimuli to begin to get degrees of freedom for testing. With a reasonable number of stimuli we could still test for unidimensionality in the face of unequal stimulus variabilities. When we come to the completely general case, allowing the correlation coefficients to vary as well, the problem is hopeless. We now have more degrees of freedom at our disposal than there are in the table. It seems reasonable then never to try to test for unidimensionality under a more general assumption than equal correlations and unequal variances for the stimuli.

6. *Conclusions*

A test of the assumptions underlying Thurstone's method of paired comparisons is developed and illustrated. The inner workings of the test and an indication of its power are provided for a special case involving three stimuli lying very close to one another. Although the method is developed and applied for Thurstone's Case V, it can be applied to any paired comparison case providing some degrees of freedom are left over after the process of estimating the spacings between the stimuli positions has been completed.

REFERENCES

1. Bartlett, M. S. The square root transformation in analysis of variance. *Supp. J. roy. stat. Soc.*, 1936, **3**, 68-78.
2. Bartlett, M. S. The use of transformations. *Biometrics Bull.*, 1947, **3**, 39-52.
3. Bliss, C. I. Plant protection, No. 12. Leningrad, 1937.
4. Fisher, R. A. On the dominance ratio. *Proc. roy. soc. Edinb.*, 1922, **42**, 321-341.
5. Mosteller, F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 1951, **16**, 3-9.
6. Mosteller, F. Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika,* 1951, **16**, 203-206.
7. Mosteller, F, and Tukey, J. W. The uses and usefulness of binomial probability paper. *J. Amer. statist. Ass.,* 1949, **44**, 174-212.
8. Statistical Research Group, Columbia University. Selected techniques of statistical analysis. New York: McGraw-Hill Book Co., 1947.

Manuscript received 9/2/50
Revised manuscript received 11/13/50