African Journal of Mathematics and Computer Science Research Vol. 5(14), pp. 274-277, December, 2012
Available online at http://www.academicjournals.org/AJMCSR
DOI: 10.5897/AJMCSR12.014
ISSN 2006-9731 ©2012 Academic Journals

Full Length Research Paper

Measuring strength of association in repeated samples

Oyeka I. C. A. and Umeh E. U.*

Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.

Accepted 19 June, 2012

This paper developed a measure of the strength of the association between populations based on ranks appropriate for the analysis of mixed effects model typed data with one observation per cell. We developed a test statistic for the proposed measure. From the result of the analysis, it was observed that the proposed method is comparable to Kendal's coefficient of concordance which assume the value zero (0) when there is perfect association and the value one (1) when there is no association whatsoever between the variables of interest.

Keywords: Association, ranks, test statistic, observation.

INTRODUCTION

When the assumptions of normality and homogeneity for the use of a parametric two way analysis of variance for data analysis are not satisfied, use of a non-parametric equivalence becomes preferable. One of the methods often used is the Friedman's two-way analysis of variance by ranks (Gibbons, 1971, 1993).

In this paper, we propose to develop a measure of the strength of the association between populations appropriate for the analysis of mixed effects model typed data with one observation per cell and to develop an alternative test statistic for the proposed measure.

THE PROPOSED MEASURE

As in Friedman's test, suppose a random sample of k assessors, judges, observers or teachers are each to observe or assess and rank each of "c" candidates, patients, conditions or situations. As in Friedman's test, this data if treated as a two-way analysis of variance would correspond to a mixed effect model without replication (Oyeka et al., 2010; Hollander and Wolfe, 1999; Siegel, 1956). This means that the data are presented in the form of k × c table with say the column corresponding to one factor with c treatments or respondents which are considered fixed and the row corresponding to a second factor with k blocks, levels or observers which are considered random and there is only "1" observation per cell. The data are therefore arranged in a table with c columns and k-rows just as for the corresponding two way analysis of variance with one observation per cell. As in the analogous analysis of variance, the null hypothesis to be tested is that the "k" judges or observers are in agreement or do not differ in their assessment of c conditions or treatments versus the alternative hypothesis that the assessors do in fact differ. Interest here is also in finding a common measure of association, agreement or concordance between the "k" assessors in their assessment of the c conditions or respondents.

To answer these questions using a non-parametric approach, we first rank the observations in each row (observer) from the smallest to the largest or from the largest to the smallest. That is, within each row (observer) the rank of 1 is assigned to the smallest (largest) value. The rank of 2 is assigned to the next smallest (largest) value and so on until the rank of c is assigned to the smallest (largest) value.

Now, let r_{ij} be the rank assigned by the i^{th} observer or assessor to the j^{th} condition, subject or object for I = 1, 2, …; j = 1, 2, …,c. Then the i^{th} row is a permutation of the numbers 1, 2, …, c in the absence of ties and the j^{th} column represents the ranks assigned to the j^{th} subject by the observers. The ranks in each column are then indicative of the agreement between observers since if the j^{th} object has the same magnitude relative to all other objects in the opinion of each of the k observers; all ranks in the j^{th} column will be the same. Thus, if the

*Corresponding author. E-mail: editus2002@yahoo.com.

null hypothesis is true, we would expect the occurrence of the ranks 1, 2, …, c to be equally likely in each column (object) across all rows (observers). This implies that we would expect the column sums of ranks to be the same under the null hypothesis. If the observed sums of column ranks are so discrepant that they are not likely to be a result of equal probabilities, then this constitutes an evidence against randomness and hence against the null hypothesis.

If however all the k observers agree perfectly in their ranking of each of the c objects, then the respective column totals $R_{.1}$, $R_{.2}$, …, $R_{.c}$ will be some permutation of the numbers 1k, 2k, …, ck. Now since the average column total is $\frac{k(c+1)}{2}$, for perfect agreement between the k observers or their ranking of the c objects, the sum of squares of deviations of column totals from the average column total, we have its maximum value, S^2_{max}which is a constant given as:

$$S^2_{max} = \sum_{j=1}^{c}\left(jk - \frac{k(c+1)}{2}\right)^2 = k^2\sum_{j=1}^{c}\left(j - \frac{(c+1)}{2}\right)^2$$

That is,

$$S^2_{max} = K^2 c\frac{(c^2-1)}{12} \qquad (1)$$

However, in general, the actual sum of squared deviations of observed column totals $R_{.j}$, from the average column total namely S^2_a is:

$$S^2_a = \sum_{j=1}^{c}\left(R.j - k\left(\frac{(c+1)}{2}\right)\right)^2$$

That is, $S^2_a = \sum_{j=1}^{c} R^2_{.j} - \frac{k^2c(c+1)^2}{4}$ (2)

Note that since S^2_{max} and S^2_a are both the sums of squares, they are both non- negative. However since k and c are both positive integer, $S^2_{max} > 0$ $(c > 1)$ but $S^2_a > 0$ and is equal to 0 if the ranking of the "c" objects by the k observers is completely at random such that $R_{.j}$ = $\frac{k(c+1)}{2}$ for all j = 1, 2, …, c if the observers are in complete agreement in their ranking of the c objects, then $S^2_a = S^2_{max}$.

Therefore, a good measure D of the strength of association between observers in their ranking of objects is the difference:

$$D = S^2_{max} - S^2_a \qquad (3)$$

Note that the smallest value that D can assume is 0 when$S^2_a = S^2_{max} = \frac{k^2c(c^2-1)}{12}$ when there is perfect association or agreement between the judges or assessors in their assessment of the subjects, treatments or conditions. The largest value D can assume is S^2_{max}, when S^2_a = 0 meaning there is independence or no association between the judges. The smaller the value of D, the stronger the association; and the larger the value of D, the stronger the disagreement between the judges.

It however seems more illuminating to have an index of association that is normed between 0 and 1with say, 0 indicating perfect association or agreement and 1 indicating independence. To achieve this objective, we divide D by S^2_{max} obtaining:

$$Q = \frac{D}{S^2_{max}} = \frac{S^2_{max} - S^2_a}{S^2_{max}}$$

That is,

$$Q = 1 - W \qquad (4)$$

$$\text{Where } W = \frac{S^2_a}{S^2_{max}} \qquad (5)$$

Is the so called Kendal's coefficient of concordance (Gibbons, 1971).

SIGNIFICANCE TEST STATISTIC FOR Q

Now the total sum of squared deviations of $r_{ij}^{'s}$ from their mean $\bar{r}$ is $S^2_t = \sum_{i=1}^{k}\sum_{j=1}^{c}(r_{ij} - \bar{r})^2$

$$= \sum_{i=1}^{k}\sum_{j=1}^{c} r^2_{ij} - k.c\left(\frac{c-1}{2}\right)^2$$

$$= \frac{kc(c+1)(2c+1)}{6} - \frac{kc(c+1)^2}{4} = \frac{kc(c^2-1)}{12}$$

Note that $S^2_{max} = k.S^2_t$ (6)

Now the total sum of squares S^2_t may be partitioned into its three component parts as:

$$S^2_t = \sum_{i=1}^{k}\sum_{j=1}^{c}(r_{ij} - \bar{r})^2 = \sum_{i=1}^{k}\sum_{j=1}^{c}\left((r_{ij} - \bar{r}_{i.} - \bar{r}_{.j} + \bar{r}) + (\bar{r}_{i.} - \bar{r}) + (\bar{r}_{.j} - \bar{r})\right)^2$$

$$= c\ {}^{k}_{i=1}\ r_{i.} - r\ {}^{2} + k.\ {}^{c}_{j=1}\ r_{.j} - r\ {}^{2} + {}^{k}_{i=1}\ {}^{c}_{j=1}\ r_{ij} - r_{i.} - r_{.j} + r\ {}^{2}$$

Where

$S^2_R = kc\left(\frac{c-1}{2} - \frac{c-1}{2}\right)^2$ = 0 is the sum of squares due to row factor namely observer;

Table 1. Weight gains (in grams) of hogs fed with certain diets. Ranks of weights shown in brackets.

Hogs (blocks)	Diets (treatment)				
	Diet 1	Diet 2	Diet 3	Diet 4	Rank total
1	1 (2)	4 (3)	8 (4)	0 (1)	10
2	2 (2)	3 (3)	13 (4)	1 (1)	10
3	10 (3)	0 (1)	11 (4)	3 (2)	10
4	12 (3)	11 (2)	13 (4)	10 (1)	10
5	1 (2)	3 (3)	10 (4)	0 (1)	10
6	10 (3)	3 (1)	11 (4)	9 (2)	10
7	4 (1)	12 (4)	10 (2)	11 (3)	10
8	10 (4)	4 (2)	5 (3)	3 (1)	10
9	10 (4)	4 (2)	9 (3)	3 (1)	10
10	14 (4)	4 (2)	7 (3)	2 (1)	10
11	3 (2)	2 (1)	4 (3)	13 (4)	10
Total $R_{.j}$	30	24	38	18	110

$$S_C^2 = k\sum_{j=1}^{c}(\bar{r}_{.j} - \bar{r})^2 = k\left(\sum_{j=1}^{c}\frac{R_{.j}^2}{K^2} - c\,\frac{(c-1)^2}{4}\right) = \sum_{j=1}^{c} R_{.j}^2 - \frac{K^2\,c(c-1)^2}{4} \quad (7)$$

is the sum of squares treatments or conditions due to column factor namely subjects, or judges where

$$R_{.j} = \sum_{i=1}^{k} r_{ij}$$

Finally,

And SSE = $S_e^2 = \sum_{i=1}^{k}\sum_{j=1}^{c}\left(r_{ij} - \bar{r}_{i.} - \bar{r}_{.j} + \bar{r}\right)^2$ is the error sum of squares.
Note that,

$$S_e^2 = \frac{S_a^2}{K} \quad (8)$$

And

$$S_e^2 = S_t^2 - S_C^2 = S_t^2 - \frac{S_a^2}{K} \quad (9)$$

It can be shown that these three sums of squares are independently distributed and that S_t^2 has a Chi-square distribution with kc-1 degrees of freedom (Hogg and Craig, 1971); S_R^2 has a Chi-square distribution with k-1 degrees of freedom; S_C^2 has a Chi-square distribution with c-1 degrees of freedom, and S_e^2 has a Chi-square distribution with $(kc-1)-(k-1+c-1)=(k-1)(c-1)$ degrees of freedom.

Hence the statistic F = $\dfrac{\frac{S_C^2}{(c-1)}}{\frac{S_e^2}{(k-1)(c-1)}} = \dfrac{(k-1)S_C^2}{S_e^2}$ (10)

has an F distribution with c-1 and $(k-1)(c-1)$ degrees of freedom. Using Equations 8 and 9 in Equation 10, we have that:

$$F = \frac{\frac{(k-1)S_a^2}{K}}{S_t^2 - \frac{S_a^2}{K}} = \frac{\frac{(k-1)S_a^2}{S_t^2 .K}}{1 - \frac{S_a^2}{S_t^2 .K}}$$

That is,

$$F = \frac{(k-1)W}{1-W}$$

has an F distribution with c-1 and $(k-1)(c-1)$ degrees of freedom which can be used to test the null hypothesis Ho: W = 0 (or Q = 1). That is, the null hypothesis of no association between judges or of independence of judges in their assessment of subjects, treatments or conditions.

ILLUSTRATIVE EXAMPLE

An experiment was conducted to determine the effects of four different types of diets on hogs. The hogs were grouped into eleven blocks in such a way that each block of four had identical environmental conditions. Each block has its four hogs assigned at random to one of the four experimental diets.

Shown in Table 1 are the weights gains (in grams) of each of the 44 hogs. The ranks assigned to the weight gains from the smallest (1) to the largest (4) within each block (hog) are shown in brackets. Interest is in determining whether hogs are different in the weight gains to the various diets and to determine the level of association between diets and weight gain by hogs.
Now from Equation 1, we have:

$$S_{max}^2 = (11)^2(4)\frac{(4^2-1)}{12} = \frac{121 \times 4 \times 15}{12} = 605$$

And from Equation 2, we have that:

$$S_a^2 = (30)^2 + (24)^2 + (38)^2 + (18)^2 - \frac{(11)^2(4)(4+1)^2}{4} = 900 + 576 + 1444 + 324 = 3244 - \frac{121 \times 4 \times 25}{4} = 3244 - 3925 = 219; S_a^2 = 219$$

Hence from Equation 3, we have that:

D = 605 - 219 = 386

The relatively smaller value of D compared with S^2_{max} would seem to suggest the existence of an association.

However, from Equation 4

Q = 1 - $\frac{219}{605}$ = 1 - 0.362 = 0.638

Hence from Equation 12; the test statistic for testing Ho: Q = 1.0 is:

F = $(11-1)\frac{(1-0.638)}{0.638} = \frac{10 \times 0.362}{0.638} = \frac{3.62}{0.638}$ = 2.3096

which has an F distribution with 3 and 30 degrees of freedom. At α = 0.05, $F_{0.95,3,30}$ = 2.92.

Since the calculated F = 2.310 < 2.92 = F tabulated, we fail to reject the null hypothesis Ho: Q = 1.0(W = 0). That is the null, hypothesis of independence or no association between hogs and types of diets. We may therefore conclude that hogs differ significantly in their response to the four types of diet. Furthermore, since Q is relatively large (Q = 63.8%), we may conclude that the association between hogs and types of diet is small. Note that for the present data, the Kendal's coefficient of concordance from Equation 5, W = $\frac{219}{605}$= 0.362 = 1 - 0.638 = 1 - Q.

Conclusion

This paper developed a measure of the strength of the association between populations based on ranks appropriate for the analysis of mixed effects model typed data with one observation per cell. We also developed a test statistic for the proposed measure.

From the aforementioned result, the proposed method is shown to be comparable to Kendal's Coefficient of Concordance and assume the value zero (0) when there is perfect association and the value one (1) when there is no association whatsoever between the variables of interest. The method is illustrated with some data.

REFERENCES

Gibbons JD (1971). Non- Parametric Statistical Inference. McGraw Hill, New York. pp. 246-250.

Gibbons JD (1993). Non- Parametric Statistical. An Introduction; Newbury Park: Sage Publication. pp. 180-220.

Hogg RV, Craig AT (1971). Introduction to Mathematical Statistics. New York, Macmillan Pub. Coy., Inc. pp. 305-340.

Hollander M, Wolfe DA (1999). Non-Parametric Statistical Methods (2nd Edition). Wiley Inter Science, New York. pp. 79-130.

Oyeka CA, Ebuh GU, Nwankwo CC, Obiora-Ilouno H, Ibeakuzie PO, Utazi C (2010). A Statistical Comparison of Test Scores: A Non-Parametric Approach. J. Math. Sci. 21(1):77-87.

Siegel S (1956). Non-Parametric Statistics for the Behavioural Sciences. McGraw-Hill, Kogakusha, Ltd. Tokyo. pp. 166-172.