IOSR Journal of Mathematics (IOSR-JM)
ISSN: 2278-5728. Volume 3, Issue 5 (Sep-Oct. 2012), PP 05-08
www.iosrjournals.org

Measure and Significance of Association between K Populations: A Non- Parametric Method

[1]Oyeka, I.C., [2]Umeh, E. U.

Abstract: *This paper presents a non parametric measure of association between k populations, and a method of testing for its significance. Analysis of variance technique is employed to developa test statistic for the measure of the association. An illustrative example is provided and the method compares equally well with the Friedman's two way analysis of variance by rank.*

I. Introduction:

When assumptions of normality and homogeneity for the use of 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 by ranks (Gibbons, 1971, Scheaffer and McClave, 1982, Gerald and Warrack, 2003, Zar, 1999, Legendre, 2005,and Sheskin, 1997).

In this paper, we propose to develop a measure of association between populations appropriate for analysis of variance by ranks and to develop an alternative test statistic for the proposed measure.

II. 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 these data if treated as a two-way analysis of variance would correspond to a mixed effects model without replication (Oyeka, 2009). This means that the data are presented in the form of a kxc table with say, the column corresponding to one factor with c treatments or respondents which are considered fixed and the row corresponding to a seccond factor with k blocks, levels or observers which are random and there are only one 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 assessors are in agreement or do not differ in their assessment of the c conditions or treatments versus the alternative hypothesis that the assessors do not 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 observation in each row (observer) from 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 or 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 largest (smallest) value.

Now let r_{ij} be the rank assigned by the ith observer or assessor to the jth condition, subject, or object, for $i = 1, 2, \dots, k, j = 1, 2, \dots, c.$ Then the $ith\ row\ is\ a\ permutation$ of the number 1, 2, …, c, and the jth column represents the ranks assigned to the jth subject by the observers. The ranks in each column are then indicative of the agreement between observers since if the jth object has the magnitude relative to all other objects in the opinion of each of the "k" observers, all ranks in the jth column will be the same. Thus if the 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 as a result of equal probabilities, then this constitutes an evidence against randomness and 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 $R_1, R_2, , \dots, R_c,$ will be some permutation of the numbers $1k, 2k, \dots, ck.$

Now since the average column total is $k\left(\frac{c+1}{2}\right),$ for perfect agreement between the k observers in their ranking of the "c" objects, the sum of squares of deviations of column totals from the average column total,S_{max}^2 will have maximum value and a constant given as:

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

That is $S_{max}^2 = k^2 c\frac{(c^2-1)}{12}$ …….. 1

However in general, the actual sum of squared deviations of observed column totals from the average total,

$$S_{ob}^2 = \sum_{j=1}^{c} \left\{ Rj - k\frac{(c+1)}{2} \right\}^2$$

That is $S_{ob}^2 = \sum_{j=1}^{c} R_j^2 - k^2 c \frac{(c+1)^2}{4}$ … …… 2

Note that since $S_{max}^2 \; and \; S_{ob}^2$ are both sums of squares, they are non negative. However since k and c are both positive integers, $S_{max}^2 > 0, (c > 1) \; but \; S_{ob}^2 \geq 0$ and is equal to 0 if the ranking of the "c" objects by the k observers are completely at random such that $R_j = \frac{k(c+1),}{2}$, for all $j = 1, 2, \ldots, c$. If the observers are in agreement in their ranking of the "c" objects, then $S_{ob}^2 = S_{max}^2$ hence a good measure W, of agreement between observers in their ranking of the objects is the ratio of these two sums of squares. That is

$W = \frac{S_{ob}^2}{S_{max}^2}$ …… 3

This is similar to Kendall coefficient of concordance (Gibbon,1971), and hence to Friedman's two- way analysis of variance without replication by ranks. Kendall's coefficient of concordance .and Friedman's two – way analysis of variance are so closely related that they address hypothesis concerning the same data table and use the same χ^2 statistic for testing (Legendry,2005). W ranges between 0 and 1 with 1 designating perfect concordance and 0 indicating no agreement or independence of populations. Usually $0 < W < 1, in \; general.$

Test Statistic for W

We now proceed to develop a test statistic for W, using analysis of variance technique. The total sum of squared deviations of assigned ranks r_{ij} from the mean rank, $\bar{r} = \frac{c+1}{2}, is$

$$SS_{total} = S_t^2 = \sum_{i=1}^{k}\sum_{j=1}^{c}\left[r_{ij} - \bar{r}\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{c}\left[r_{ij} - \frac{(c+1)}{2}\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{c} r_{ij}^2 - \frac{kc(c+1)^2}{4}$$

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

That is

$SS_{total} = S_t^2 = \frac{kc(c^2-1)}{12}$ …… 4

Note from equations 1and 4 that

$S_{max}^2 = kS_t^2$ …… 5

The total sum of squares $SS_{total} = S_t^2 \; can$ be partitioned into three sums of squares that can be shown to be independent (Hogg and Craig, 1971).

Thus $SS_{total} = S_t^2 = \sum_{i=1}^{k}\sum_{j=1}^{c}\left(r_{ij} - \bar{r}\right)^2$

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

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

Now $\sum_{i=1}^{k}\sum_{j=1}^{c}\left(r_{ij} - \bar{r}_{i.} - \bar{r}_{.j} + \bar{r}\right)^2 \; is \; the \; error \; sum \; of \; squared \; deviation \; , SS_E = SS_e$

$c\sum_{i=1}^{k}(\bar{r}_{i.} - \bar{r})^2 = ck\left[\frac{(c+1)}{2} - \frac{(c+1)}{2}\right]^2 = 0$is sum of the squared deviations due to row or observers.

Finally, $k\sum_{j=1}^{c}\left(\bar{r}_{.j} - \bar{r}\right)^2 = k\left[\sum_{j=1}^{c}\frac{R_j^2}{R^2} - \frac{c(c+1)^2}{4}\right] = \frac{\left\{\sum_{j=1}^{c} R_j^2 - \frac{k^2 c(c+1)^2}{4}\right\}}{k}$ is the sum of squared deviations due to column (object),$SS_c = S_c^2$

That is

$SS_c = S_c^2 = \sum_{j=1}^{c} R_{.j}^2 - \frac{\frac{k^2 c(c+1)^2}{4}}{k}$ …… 6

In other words, $S_c^2 = \frac{S_{ob}^2}{k}$ ……... 7

Therefore, $S_t^2 = \frac{S_{ob}^2}{k} + S_e^2$

8

Now for sufficiently large values of k and c, it is known that the observer or row sum of squares SS_R which is zero has a chi-square distribution with k-1 degrees of freedom, the object or column sum of squares, $SS_c = \frac{S_{ob}^2}{k}$ has a chi-square distribution with c-1 degrees of freedom and the sum of squares error, $SS_E = SS_e$ has a chi-square distribution with $(k-1)(c-1)$ degrees of freedom (Hogg and Craig, 1971). Hence under H_0:

$$F = \frac{SS_c/_{(c-1)}}{SS_e/_{(k-1)(c-1)}} = \frac{(k-1)S_{ob}^2/_k}{S_e^2} \quad ---$$

9

Has an F- distribution with $(c-1)$and $(k-1)(c-1)$ $degrees\ of\ freedom$ or from equation 8, we have that:

$$F = \frac{(k-1)S_{ob}^2/_k}{S_t^2 - \frac{S_{ob}^2}{k}} \quad ---$$

10

has an F-distribution with $(c-1)$and $(k-1)(c-1)$ $degrees\ of\ freedom.$

Using equation 5 in 10, we have that

$$F = \frac{(k-1)S_{ob}^2}{kS_t^2 - S_{ob}^2} = \frac{(k-1)S_{ob}^2}{S_{max}^2 - S_{ob}^2} \quad ---$$

11

Dividing through equation 11 by S_{max}^2 and noting from equation 3 that

$W = \frac{S_{ob}^2}{S_{max}^2}$we have the test statistic

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

12

Which has an F- distribution with $(c-1)$and $(k-1)(c-1)degrees\ of\ freedom$ which can be used to test our H_0 about W. $H_0\ is\ to\ rejected\ at\ \alpha\ level\ of\ significance\ if$

$F \geq F_{1-}\alpha,(c-1),(k-1)(c-1)$ ---

13

Accept otherwise.

ILLUSTRATIVE EXAMPLE

The percent reduction in heart beat of a random sample of 15 bats of ertain species after the administration of three different dose levels of a certain drug is presented in Table 1.

Table 1. Bats and Dose levels of the Drug

Bat No.				
	A	B	C	Total
1	2	3	1	6
2	2	1	3	6
3	1	2	3	6
4	1	2	3	6
5	2	1	3	6
6	2	3	1	6
7	3	1.5	1.5	6
8	1	3	2	6
9	3	2	1	6
10	1	2.5	2.5	6
11	3	1.5	1.5	6
12	3	1	2	6
13	1	2.5	2.5	6
14	3	1	2	6
15	1.5	1.5	3	6
Total	29.5	28.5	32	90

,

Source: Exercises at the end chapter 14 Question 14.12 (Oyeka, 2009). Interest is in testing at 0.01 level of significance, the null hypothesis of no difference in responses between the three dose levels A, B, C. Or symbolically the null hypothesis of interest is stated thus:

H_0: The locations of all k populations are the same

H_1: At least two populations differ

Then we obtained from computations as follows:

from Equation 4

$$S_t^2 = \frac{15(3)(8)}{12} = 30$$

FromEquation 1, $S_{max}^2 = \frac{15^2(3)(8)}{12} = 45.0$

From equation 2, $S_{ob}^2 = (29.5)^2 + (28.5)^2 + (32)^2 - \frac{15^2 x 3 x 4^2}{4} = 6.5$

From equation 3, W $W = \frac{6.5}{45.0} = 0.014$

From Equation 7, $S_c^2 = \frac{6.5}{15} = 0.433$

S_e^2 = 30-0.433 = 29.567

And from Equation 12, we have

$$F = \frac{14(0.014)}{1 - 0.014} = 0.199 \quad (pvalue = 0.8207)$$

But $F_{0.99,2,28} = 5.45$

Since $F = 0.199 < 5.45 = F_{0.99,2,28}$, we accept H_0 and conclude that there is no significance difference in responses of the bats to three dose levels of the drug.

Friedman's Two- Way Analysis of Variance by Rank Method

Table 2: Bats, Dose Levels of Drugs and their Ranks

Bat No	Dose Levels			Ranks			
	A	B	C	Rank (A)	Rank (B)	Rank ©	Total
1	5	6	3	2	3	1	6
2	6	4	8	2	1	3	6
3	2	3	8	1	2	3	6
4	2	5	7	1	2	3	6
5	3	2	4	2	1	3	6
6	4	5	3	2	3	1	6
7	12	7	7	3	1.5	1.5	6
8	6	12	7	1	3	2	6
9	7	5	3	3	2	1	6
10	3	4	4	1	2.5	2.5	6
11	4	3	3	3	1.5	1.5	6
12	8	6	7	3	1	2	6
13	2	7	7	1	2.5	2.5	6
14	13	7	8	3	1	2	6
15	5	5	10	1.5	1.5	3	6
Total				29.5	28.5	32	90

To test the null hypothesis that the locations of all k populations are the same against the alternative that at least two populations' locations differ, Friedman's F- ratio (Fr) test statistic is

$Fr = \frac{12}{kc(c+1)} \sum_{j=1}^{c} R_{.j}^2 - 3k(c+1)$ ___ 14

and rejection is given by

$$Fr = \frac{12}{15(3)(4)}(29.5^2 + 28.5^2 + 32) - 3(15)(4)$$
$$= 0.433 \ (p\ Value = 0.8057$$

and$\chi_{0.99,2}^2 = 9.21$

Since $Fr = 0.433 < 9.21 = \chi_{0.99,2}^2$, we accept $H_0$and conclude that the responses of the bats to the three dose levels of the drug do not differ.

III. Conclusion

Both the proposed method and Friedman's two way analysis of variance by rank method not only accepted H_0 at 1% significant levels, but also their p-values were almost the same (0.8207, and 0.8057 respectively). Thus one can conclude that the proposed test statistic is as good as the Friedman's test statistic in this case.

Reference

[1] Gerald, K. and Warrack , B. (2003): Statistics for Management and Economics, Curt Hinrichs, USA

[2] Gibbons, J.D. (1971): Non Parametric Statistical Inference, McGraw Hill, New York

[3] Legendre, P. (2005): Species Associations: The Kendall Coefficient Revisited, American Statistical Association and International Biometric Society, Journal of Agricultural, Biological, and Environmental Statistics, Vol 10, 2

[4] Oyeka, C.A. (2009): An Introduction to Applied Statistical Methods (5theds), Nobern Avocation Pub. Coy., Enugu

[5] Scheaffer, R.L. and McClave, J.T. (1982): Statistics for Engineers, PWS, Publishers, USA

[6] Sheskin, D.J (1997): Handbook of Parametric Statistical Procedures, CRC Press, Inc, Boca Raton, New York

[7] Zar, J.H (1999), Biostatistical Analysis (4theds), Prentice Hall, Upper Saddle River, New Jersey